electronics

MDPI

Article

Applying the Lombard Effect to Speech-in-Noise Communication

Gražina Korvel [1,*], Krzysztof Kąkol [2], Povilas Treigys [1] and Bożena Kostek [3,*]

[1] Institute of Data Science and Digital Technologies, Vilnius University, LT-08412 Vilnius, Lithuania; povilas.treigys@mif.vu.lt
[2] PGS Software, 50-086 Wrocław, Poland; krzysztofkakol@gmail.com
[3] Audio Acoustics Laboratory, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, 80-233 Gdansk, Poland
* Correspondence: grazina.korvel@mif.vu.lt (G.K.); bokostek@audioakustyka.org (B.K.)

Abstract: This study explored how the Lombard effect, a natural or artificial increase in speech loudness in noisy environments, can improve speech-in-noise communication. This study consisted of several experiments that measured the impact of different types of noise on synthesizing the Lombard effect. The main steps were as follows: first, a dataset of speech samples with and without the Lombard effect was collected in a controlled setting; then, the frequency changes in the speech signals were detected using the McAulay and Quartieri algorithm based on a 2D speech representation; next, an average formant track error was computed as a metric to evaluate the quality of the speech signals in noise. Three image assessment methods, namely the SSIM (Structural SIMilarity) index, RMSE (Root Mean Square Error), and dHash (Difference Hash) were used for this purpose. Furthermore, this study analyzed various spectral features of the speech signals in relation to the Lombard effect and the noise types. Finally, this study proposed a method for automatic noise profiling and applied pitch modifications to neutral speech signals according to the profile and the frequency change patterns. This study used an overlap-add synthesis in the STRAIGHT vocoder to generate the synthesized speech.

Keywords: Lombard effect; noise background; Structural SIMilarity (SSIM) index; RMSE (Root Mean Square Error); dHash (Difference Hash)

Citation: Korvel, G.; Kąkol, K.; Treigys, P.; Kostek, B. Applying the Lombard Effect to Speech-in-Noise Communication. *Electronics* **2023**, *12*, 4933. https://doi.org/10.3390/electronics12244933

Academic Editors: Athanasios Koutras and Chrisoula Alexandraki

Received: 5 November 2023
Revised: 1 December 2023
Accepted: 5 December 2023
Published: 8 December 2023

1. Introduction

This study aims to improve speech-in-noise communication using the Lombard effect (LE), which is a natural or synthesized adaptation of speech loudness and style in response to noise [1]. The Lombard effect is a phenomenon that occurs when people speak more loudly and change their speech style to be heard and understood in noisy situations [2]. The Lombard effect is named after Étienne Lombard, a French doctor who first described it in 1909 [3]. LE involves not only increasing the volume of the voice but also other adjustments, such as raising the pitch, lengthening the syllables, and shifting the energy of the speech to higher frequencies [2,3]. These changes help to improve the clarity and intelligibility of the speech in noise. It should be noted that LE is involuntary and reflexive, meaning that people do not consciously control it. It is influenced by both the level and the type of noise, as well as the speaker's hearing feedback [4]. The Lombard effect has implications in various fields and applications, such as speech recognition, hearing aids, voice synthesis, and acoustic design [4–6].

Recent research further supports the benefits of the Lombard effect in improving speech intelligibility. Hansen et al. [7] investigated how the artificially induced Lombard effect, created by incorporating three alternative modification techniques based on (i) durational modification, (ii) temporal amplification of highly intelligible segments, and (iii) spectral mismatch filtering, significantly improved speech intelligibility for cochlear implant users in noisy environments. Similarly, Vlaj and Kacic [8] investigated the impact

of the Lombard effect on speech recognition, highlighting its potential to improve communication in challenging acoustic environments. Kang et al. [9] took a different approach by optimizing a real-time wavelet-based algorithm that adjusts the sub-band increments of speech frequencies. Their work has demonstrated a significant improvement in speech intelligibility even in the presence of background noise, confirming the practical applicability of methods inspired by the Lombard effect.

This study has two main objectives: first, to understand how noise affects the speech features that are related to LE; and second, to train a noise recognition model that can automatically generate LE speech when noise is detected and identified. So, to build a human-centric system with ambient intelligence to generate LE speech for better intelligibility, first, it needs to learn about noise interference on speech characteristics. Second, to enable the system to generate Lombard speech automatically when noise is detected and correctly labeled, the interference sound recognition model should be trained on speech with this phenomenon present; however, this task is challenging because there is a lack of data on LE speech, which is needed for deep-learning models [10]. Therefore, this study proposes to use a text-to-speech (TTS) system with a suitable vocoder to synthesize LE speech, which can perform better than conventional methods in low SNR (signal-to-noise ratio) conditions. This approach is based on the studies by [10,11], which confirmed the effectiveness of TTS and vocoders for LE speech synthesis.

LE is a phenomenon that has attracted a lot of attention from various researchers who are aiming to enhance the performance of automatic speech recognition systems in noisy environments [12] and improve speech intelligibility, by transforming the speaking style from normal (neutral) to Lombard speech [13,14], or increase speech intelligibility in cochlear implant patients [10]. Another potential application of LE is to enable speech synthesizers to adapt to noisy conditions, which positively impacts intelligibility gain [15–19]. Despite this, when one refers to the recognition of real-life speech in noise, and especially when noise profiling is a necessary step to process the speech signal correctly, the progress in this area is below expectation, even though some preliminary studies have shown promising results in this direction [20–22]. It should, however, be remembered that while humans are good at understanding speech in noisy conditions, including cocktail-party environments, one would expect that an algorithmic approach can reproduce such a possibility to some extent and produce substantial improvements in speech-in-noise intelligibility. However, this is not free from limitations, resulting in unsatisfactory outcomes.

One of the challenges in this research was to collect more data on Lombard speech for deep models. Therefore, the proposed method was to automatically detect the presence of LE in speech from the Internet and use it to train deep networks. To achieve this, the differences between clean speech with LE and Lombard speech in noise were analyzed. Specifically, the rapidly varying regions of speech, such as the transitions between voiced and unvoiced segments, were investigated. The transitions between voiced and unvoiced segments are critical for speech intelligibility. In noisy environments, where Lombard speech typically occurs, these transitions help to distinguish speech sounds. Understanding how these transitions are modified under the Lombard effect contributes to improving speech intelligibility in noise. These regions can be identified by estimating the frequency tracks and their spectral energy peaks. The number and location of the peaks are relevant for this task. This analysis covers different types of noise and SNR levels.

This article is built upon the paper presented by the authors at the ISMIS conference [19]. However, a thorough analysis of tracking changes in 2D speech representations, i.e., spectrograms, mel spectrograms, chromagrams, and MFCC-grams, by introducing the noise of various SNRs, is included, as well as additional of noise types to the investigation, comparing with the original conference paper.

2. Materials and Methods

To investigate the variation in frequency characteristics of Lombard speech at different noise distortions, the speech signals were converted to 2D representations, i.e., spectro-

grams, mel spectrograms, chromagrams, and MFCC-grams. Based on such a representation, frequency tracks were extracted employing the McAulay and Quartieri [23] algorithm. An overlap-add synthesis using a minimum-phase impulse response with group delay manipulations implemented in the STRAIGHT vocoder was used for speech synthesis. The block diagram of the experimental setup is presented in Figure 1. As seen in Figure 1, an evaluation applied to monitor frequency changes in the presence of noise was performed using several measures. After the frequency track changes were assessed, automatic noise profiling, followed by pitch modifications of natural speech signals depending on the profiling result and frequency change trends, was obtained. Perceptual evaluation was applied to check the quality of the synthesized speech, into which LE was incorporated according to noise profiling.

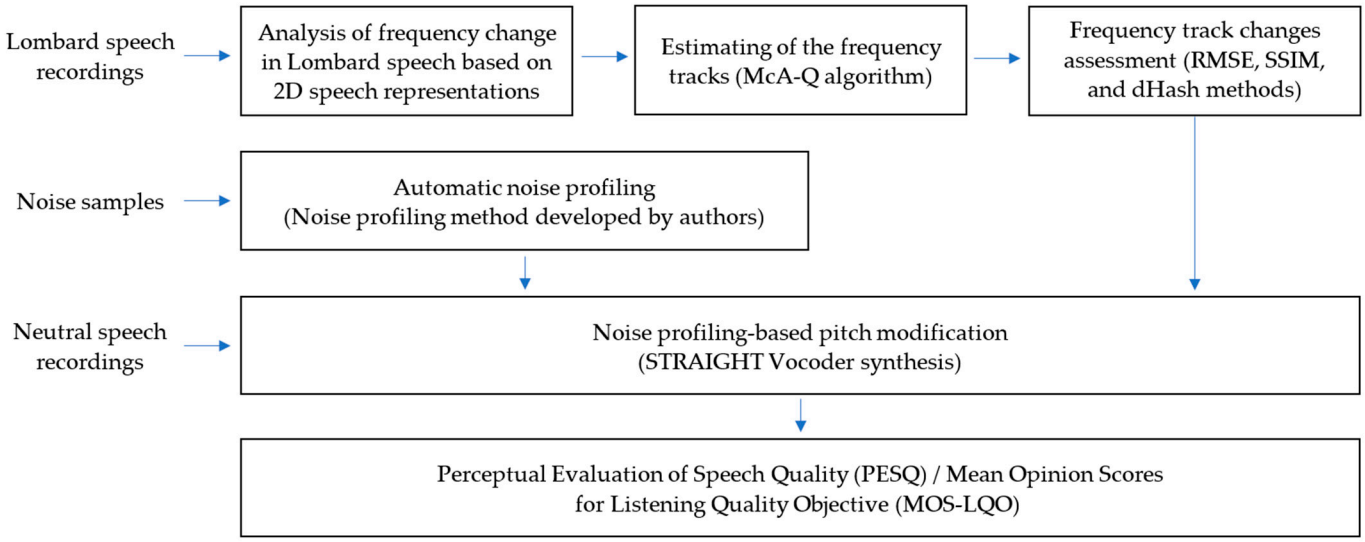


Figure 1. Block diagram of the experiment.

2.1. Analysis of Frequency Change in Lombard Speech

To detect the frequency changes at each time point, the time–frequency signal features were converted to the following 2D representations: spectrograms, mel spectrograms, chromagrams, and MFCC-grams. The process of 2D speech signal representation creation consists of the calculation of the discrete Fourier transform of each short-time frame of speech signal:

$$X_l(k) = \sum_{n=0}^{N-1} x_l(n) w(n) e^{\frac{-2\pi jkn}{N}} \quad (1)$$

where $X_l(k)$ are Fourier transform coefficients ($k = 0, \ldots, N_{FT} - 1, N_{FT}$ is the number of Fourier transform coefficients), $x_l(n)$ is the sample of lth short-time frame of signal ($l = 0, \ldots, L - 1$, and L denotes the number of short-time frames), N is the length of the signal, $w(n) = 0.54 - 0.46\text{con}(2\pi n / N - 1)$ is the Hamming window function, and j is the imaginary unit.

(1) Spectrogram

To generate a spectrogram, the Fourier transform coefficients are collected together, and a spectrogram image is built up.

(2) Mel spectrogram

The mel spectrogram is a scaled power spectrogram created using a filter bank over a specific frequency range. The relationship between the mel scale and the Hertz scale is defined by the following formula:

$$Mel(f) = 2595 \log\left(1 + \frac{f}{700}\right) \tag{2}$$

where f is a given frequency in Hertz.

(3) Chromagram

The chromagram representation projects the entire musical spectrum into 12 bins that correspond to the 12 semitones of an octave. This results in an observation that every pitch might be represented by two factors: tone height and chroma. Tone height is represented by the octave number, while the chroma is the number of pitches inside the octave (0 to 11)—just like sounds in a chromatic scale (C–C#–D–D#–. . .–B).

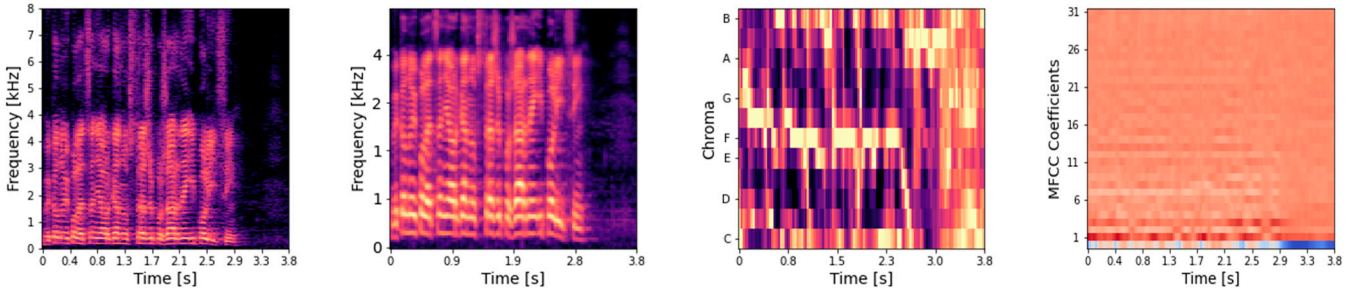
(4) MFCC-gram

MFCCs, short for Mel-frequency Cepstral Coefficients, provide a condensed version of the mel spectrogram. The log magnitude of the mel spectrum is first computed, and then the Discrete Cosine Transformation (DCT) is applied to obtain MFCCs. The mathematical formula for MFCCs can be expressed as follows:

$$c_n = \sum_{i=0}^{M-1} m_i \cos\left(\frac{\pi n(i + 1/2)}{M}\right) \tag{3}$$

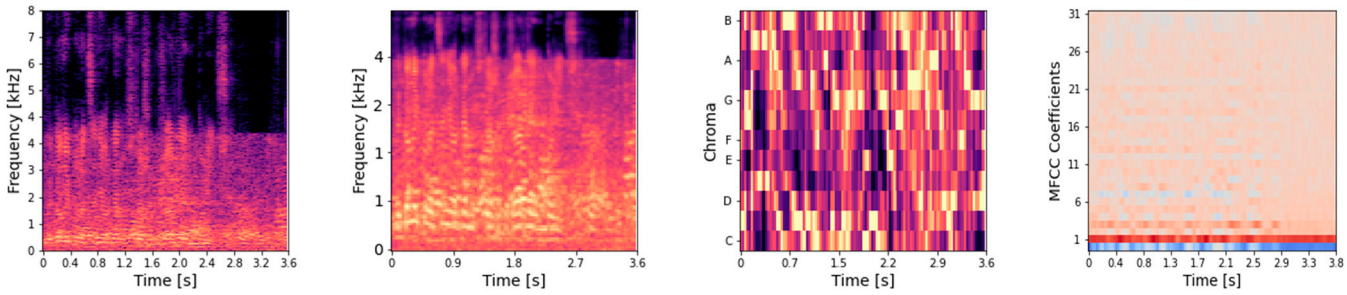
where m_i are the log filter bank amplitudes in i-th mel filter bank, M is the number of filters in the mel filter bank, and n refers to the order of the cepstral coefficient being calculated ($n = 0, \ldots, M - 1$).

These feature space representations are visualized in Figures 2 and 3, where the Lombard speech excerpt and the same speech fragment with added nonstationary street noise at 0 dB SNR are displayed, respectively. For this analysis, the spectrogram representation was generated using Hamming windows of size 512. This window size gives a smoothed Fourier spectrum. At the same time, the frequency resolution is sufficient for frequency tracking. An overlap of 256 is used to avoid losing information due to the window operation.



(**a**) (**b**) (**c**) (**d**)

Figure 2. Lombard speech signal representations: (**a**) spectrogram, (**b**) mel spectrogram, (**c**) chromagram, (**d**) MFCC-gram. The colors represent the intensity of a particular features at specific points in time, with brighter colors indicating higher intensities and darker colors indicating lower intensities.



(**a**) (**b**) (**c**) (**d**)

Figure 3. Lombard speech signal representations with additive nonstationary restaurant noise at 0 dB SNR: (**a**) spectrogram, (**b**) mel spectrogram, (**c**) chromagram, (**d**) MFCC-gram. The colors represent the intensity of a particular features at specific points in time, with brighter colors indicating higher intensities and darker colors indicating lower intensities.

Upon analyzing the 2D signal representations, it became apparent that the spectrogram was the most effective in illustrating how the signal's frequency content varies over time. The spectrogram provides a detailed view of both the time and frequency domains, enabling us to examine changes in the frequency range with good detail.

Changes in frequency can also be well observed in the mel spectrogram. On the other hand, the mel scale is specifically designed to reflect how the human auditory system perceives sound. It emphasizes lower frequencies and is less sensitive to differences in higher frequencies. Hence, a spectrogram that provides a more straightforward representation was chosen.

2.2. Estimation of Frequency Tracks

Several methods for tracking frequency tracks and their variations have been proposed in the literature [24,25]. In this study, the classic algorithm by McAulay and Quartieri (McA-Q) was adopted [23]. This algorithm detects frequency tracks in spectrograms by finding the local maxima of the spectrogram in each short-time frame, l. These maxima are called peaks. The peaks, along with their amplitudes and frequencies, are then fed to the tracking algorithm, which aims to eliminate partial trajectories. The McA-Q algorithm performs peak matching between consecutive frames. The algorithmic form of the peak matching process between frames l and $l+1$ is presented in Appendix A. The process of determining frequency tracks in a speech signal is performed based on a spectrogram, a visual representation of the distribution of signal acoustic energy across frequencies and over time. The color darkness reflects the signal intensity.

The matching of each spectrum peak in frame l to the peaks in frame $l+1$ consists of 3 main steps. In the first step, for each frequency ω_n^l in frame l, a search is done for a frequency ω_m^{l+1} in frame $l+1$, which is the nearest to this frequency and whose absolute distance is less than the threshold (i.e., Δ). In the second step, it is checked if the frequency ω_m^{l+1} has no better match to the unmatched frequencies of frame l. If this condition is satisfied, then the frequencies are matched, and their amplitudes are interpolated between the frames. Otherwise, the adjacent remaining lower frequency ω_{m-1}^{l+1} (if such exists) is tested. In the last step, for the remaining frequencies in frame $l+1$, for which no matches were made, frequencies are created in frame l with zero amplitude, and the match is made.

The result of applying the tracker to the Lombard speech signal is shown in Figure 4, where the Lombard speech excerpt and the same Lombard speech excerpt with added nonstationary street noise at 0 dB SNR are displayed.

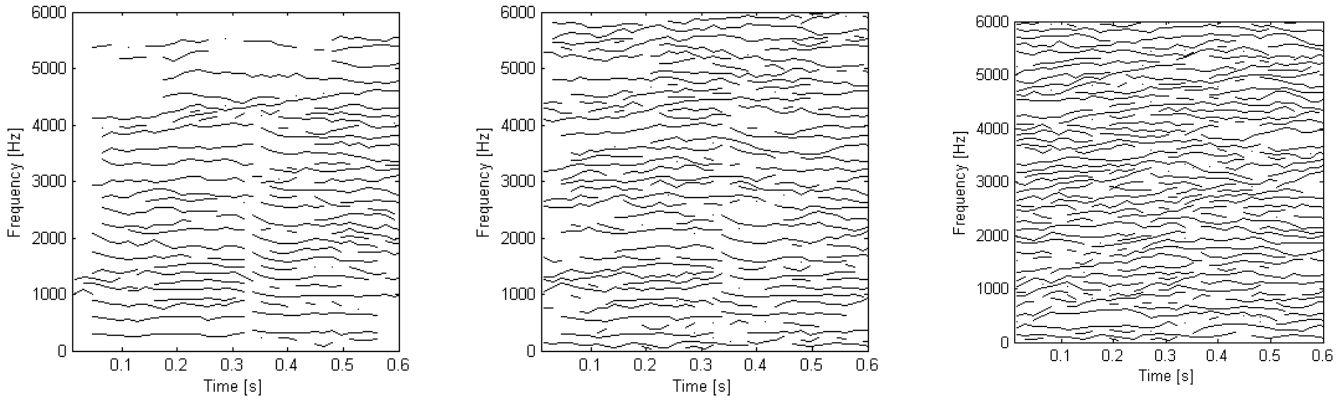


Figure 4. Lombard speech signal estimated frequency tracks (**a**) without additive noise, (**b**) with additive nonstationary street noise at 20 dB SNR, (**c**) with additive nonstationary street noise at 0 dB SNR.

A ratio of 0 dB indicates equal signal and noise levels; therefore, the degradation of formant tracks of noisy speech is visible (Figure 4). When SNR is 20 dB, the signal can generally be considered relatively clean, and frequency track tendencies compared with a signal without added noise can be perceptible. When SNR is 0 dB, it can be regarded as heavily noisy. To verify the difference in intelligibility between these three conditions, behavioral or psychophysical methods should be employed [26].

2.3. Frequency Track Changes Assessment

To evaluate frequency track changes, three different image assessment measures were used: SSIM (Structural SIMilarity) index—an image-quality assessment metric that quantifies the similarity between two images based on their structure, luminance, and contrast; RMSE (Root Mean Square Error)—a commonly used metric to measure the difference between two sets of values (including two images) based on the calculation of the square root of the average of the squared differences between corresponding values; and dHash (Difference Hash)—a technique used for image hashing, which occurs together with the Hamming distance to compare obtained hash values.

2.3.1. The Structural Similarity Index Measure

To quantify the effect of noise on speech, in terms of the Lombard effect, an average formant track error was calculated as an objective image-quality metric. Firstly, the Structural SIMilarity (SSIM) index was calculated for image-quality assessment. The SSIM index was developed by Wang et al. [27] to evaluate the quality of two images based on the perspective of image formation, i.e., the image luminance, contrast, and structural similarity. The advantages mentioned above for this method make it sensitive to changes in the image, which is very important in this study. It should also be noted that the SSIM index is widely used as the quality indicator of compared images [28,29].

Let x and y be two non-negative image signals. The structural SSIM index is calculated using the following formula [27]:

$$S(x,y) = \left[l(x,y)\right]^{\alpha} \cdot \left[c(x,y)\right]^{\beta} \cdot \left[s(x,y)\right]^{\gamma} \tag{4}$$

where $\alpha > 0$, $\beta > 0$, and $\gamma > 0$ are weights (in this research parametrized as $\alpha = \beta = \gamma = 1$), $l(x,y)$ is the luminance comparison function, $c(x,y)$ is the contrast comparison function, and $s(x,y)$ is the structure comparison function. The functions are given by:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{5}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{6}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{7}$$

where μ_x and μ_y, σ_x and σ_y, and σ_{xy} are the local means, standard deviations, and cross-covariance of the images being compared, respectively. The constants C_1, C_2, and C_3 are used to avoid instability [27]. The overall similarity measure, SSIM, is in the range of -1 to 1. A value of 1 indicates an ideal agreement between two images, while a value of -1 indicates that the given images are very different. In this research, the difference between the image of the estimated frequency tracks of the clean speech signal and that of the noisy speech signal was calculated using the SSIM index.

2.3.2. The Root Mean Square Error Measure

The Root Mean Square Error (RMSE) is a commonly used measure to evaluate the difference between two sets of values, including two images [30,31]. It is calculated as the square root of the average of the squared differences between corresponding pixel values in two images:

$$\text{RMSE} = \frac{1}{MN}\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}[A(i,j) - B(i,j)]^2} \tag{8}$$

where $A(i,j)$ and $B(i,j)$ are the pixel values at position (i,j) in reference image A and fused image B, respectively, and M and N are the width and height of the images, respectively.

A lower RMSE value indicates better image similarity. The increased value of RMSE means that the images are less similar and the degree of distortion is higher.

2.3.3. The dHash-Based Measure

Hashing is a process that uses a mathematical algorithm to transform input data into a unique output called a hash value. A hash value is a fixed-length string of characters representing the original data. This process is useful for verifying the integrity and authenticity of data, as any change in the input data will result in a different hash value. Hashing is also used for storing and retrieving data efficiently, as hash values can be used as indexes in a data structure. Some examples of hashing algorithms are MD5, SHA-1, SHA-256, and RIPEMD-160 [31–34].

Regarding images, the hashing function can generate hash codes that capture the unique characteristics of images, allowing for image comparison [35], image retrieval [36,37], and image authentication [38]. Hash functions are ideal for detecting (near-)identical photos because of their robustness against minor changes, while also minimizing the number of false-positive collisions [39].

By comparing the hash value of two images, their similarity is determined. The difference hash (dHash) algorithm, based on the calculation of the difference for each of the pixels and comparing the difference with the average differences, was used in this research. The algorithm operates as follows: the input image is first converted to greyscale and resized to a smaller size; the pixel differences are then calculated. This process is repeated for all rows, thus producing a row-wise hash. The row-wise hashes represent the final hash value.

The pixel differences are calculated as follows: if the value of the left pixel is greater than or equal to the value of the right pixel, a 1 is assigned to represent that pixel pair. Otherwise, a 0 is assigned. These values are the individual parts that make up the resulting dHash value.

To estimate the similarity between the two images, the Hamming distance is calculated based on the dHash values of the two images:

$$Hamming\ distance = Number\ of\ differing\ bits \quad (9)$$

This paper used the Python module, ImageHash, which was developed by Buchner [40]. The hash consists of 16 hexadecimal characters, and each hexadecimal character represents 4 bits; therefore, the size of this hash is 64 bits. In the case of 64-bit hashes, the Hamming distance ranges from 0 (when the hashes are identical) to 64 (when all hashes are different). According to Joshi et al., if the Hamming distance is less than 5, the images are considered similar or duplicates [39].

3. Experiments

Recordings of natural speech were obtained in a room with an acoustically treated interior that suppresses reverberation. Recordings of Lombard speech were obtained in the same studio to maintain the same acoustic conditions. To acquire the Lombard effect while speaking, the interfering noise was played back through closed headphones. Eight speakers (four males and four females) read fifteen sentences separately. The speakers were untrained, healthy, native students from the Gdansk University of Technology. Each speaker repeated a given sentence twice under a different condition.

The Lombard speech recordings were split into smaller segments, the length of which was 1 s. As a result, 2719 recordings were used in the experiment. The effect of different types of noise was investigated at varying levels of SNR, from −10 dB to 40 dB (i.e., from high to slightly distorted speech).

Two types of noise, generated and environmental, were added to the signals. Four real-life noises, babble speech (i.e., a mix of many talkers), city streets, rain, and pub recordings, were selected. These recordings were taken from the YouTube platform. High-quality recordings were selected to minimize the loss of compression artifacts. After downloading, spectral analysis was also performed on each recording. The speech and noise signal sampling rates were adjusted to 16 kHz before the test. Generated noise included two colored noises, namely pink and purple.

3.1. Results of the Influence of Noise Interference on the Frequency Tracks of Lombard Speech

The experiment was designed to measure the influence of noise interference on the frequency tracks of Lombard speech. The obtained SSIM index values, RMSE values, and dHash-based values indicating the correspondence between the shape of a speech signal with LE and its noisy version in different SNR conditions are contained in Tables 1–3, respectively.

Table 1. The SSIM index values for Lombard speech recordings (the SSIM index ranges from 0 to 1, where a higher value indicates greater similarity, while a value of 0 indicates dissimilar images).

Noise Type		−10 dB	0 dB	10 dB	20 dB	30 dB	40 dB
Purple noise	Mean	0.3439	0.347	0.3548	0.3875	0.5461	0.7392
	STD	0.0006	0.0007	0.0008	0.0014	0.0034	0.0043
Pink noise	Mean	0.343	0.3444	0.3621	0.5258	0.6958	0.8095
	STD	0.0006	0.0007	0.0011	0.0031	0.0041	0.0037
Pub noise	Mean	0.412	0.428	0.514	0.634	0.767	0.862
	STD	0.001	0.001	0.002	0.003	0.003	0.002
City street noise	Mean	0.3733	0.3789	0.4198	0.5379	0.6985	0.8188
	STD	0.0007	0.0007	0.0013	0.0025	0.0033	0.0030

Table 1. *Cont.*

Noise Type		−10 dB	0 dB	10 dB	20 dB	30 dB	40 dB
Babble speech noise	Mean	0.5214	0.5618	0.6610	0.7656	0.8547	0.9146
	STD	0.0015	0.0019	0.0026	0.0027	0.0021	0.0016
Rain noise	Mean	0.3672	0.3701	0.3843	0.5214	0.6959	0.8202
	STD	0.0006	0.0007	0.0009	0.0026	0.0035	0.0031

Table 2. The RMSE values for Lombard speech recordings (the RMSE measure does not have a predefined range. The RMSE score is a non-negative real number, and 0 means that the images are precisely the same).

Noise Type		−10 dB	0 dB	10 dB	20 dB	30 dB	40 dB
Purple noise	Mean	3.8962	3.8910	3.8833	3.8469	3.3810	2.7931
	STD	0.0350	0.0349	0.0346	0.0346	0.0393	0.0431
Pink noise	Mean	3.5239	3.5156	3.5003	3.4697	3.1240	2.5524
	STD	0.0434	0.0432	0.0427	0.0420	0.0459	0.0477
Pub noise	Mean	3.8153	3.8110	3.8059	3.7779	3.3486	2.7671
	STD	0.0363	0.0363	0.0360	0.0362	0.0410	0.0430
City street noise	Mean	3.7879	3.7837	3.7796	3.7679	3.3289	2.7807
	STD	0.0421	0.0419	0.0417	0.0416	0.0480	0.0517
Babble speech noise	Mean	3.7813	3.7708	3.7541	3.7139	3.2991	2.6972
	STD	0.0310	0.0309	0.0306	0.0308	0.0350	0.0377
Rain noise	Mean	3.6019	3.5930	3.5773	3.5498	3.1656	2.6105
	STD	0.0337	0.0335	0.0332	0.0328	0.0361	0.0380

Table 3. The dHash-based values for Lombard speech recordings (the Hamming distance ranges from 0 to 64, where a lower value indicates more remarkable similarity, while a value of 64 indicates dissimilar images).

Noise Type		−10 dB	0 dB	10 dB	20 dB	30 dB	40 dB
Purple noise	Mean	25.172	24.666	24.185	23.267	19.584	11.238
	STD	0.161	0.166	0.158	0.171	0.224	0.268
Pink noise	Mean	25.449	25.110	24.637	20.627	14.746	10.445
	STD	0.168	0.170	0.175	0.211	0.246	0.250
Pub noise	Mean	25.391	24.357	22.633	18.433	13.128	9.674
	STD	0.165	0.172	0.180	0.214	0.244	0.233
City street noise	Mean	25.564	25.008	23.728	20.769	15.177	10.609
	STD	0.173	0.174	0.177	0.197	0.238	0.239
Babble speech noise	Mean	23.934	21.729	17.923	13.554	10.098	7.284
	STD	0.169	0.200	0.230	0.239	0.217	0.189
Rain noise	Mean	25.480	25.012	24.398	21.017	15.013	10.657
	STD	0.169	0.167	0.166	0.198	0.236	0.242

To compare the results of the three measures, the results were normalized to the interval [0, 1]. For RMSE and dHash-based scores, the results were further transformed. In the transformed version, the biggest value from the original scores was represented as 0, and the smallest value was defined as 1, with the other values scaled accordingly between

0 and 1. The STD for each normalized mean value was calculated using the following formula:

$$Normalized\_STD = \frac{Range\_of\_Normalized\_Mean * Original\_STD}{Range\_of\_Original\_Mean} \quad (10)$$

Range_of_Normalized_Mean and *Range_of_Original_Mean* refers to the difference between the maximum and minimum values of the normalized mean values and original mean values, respectively.

A graphical representation of the results obtained is given in Figures 5–7, where the normalized standard deviation is shown as a vertical error bar.

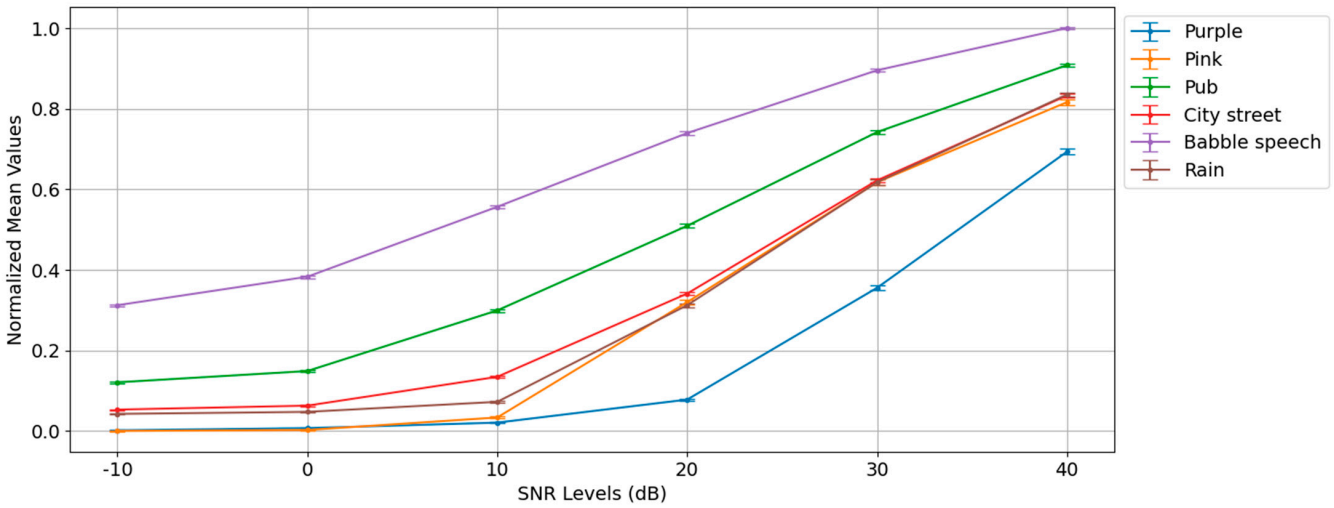


Figure 5. The mean and standard deviation of the normalized SSIM index values for Lombard speech recordings.

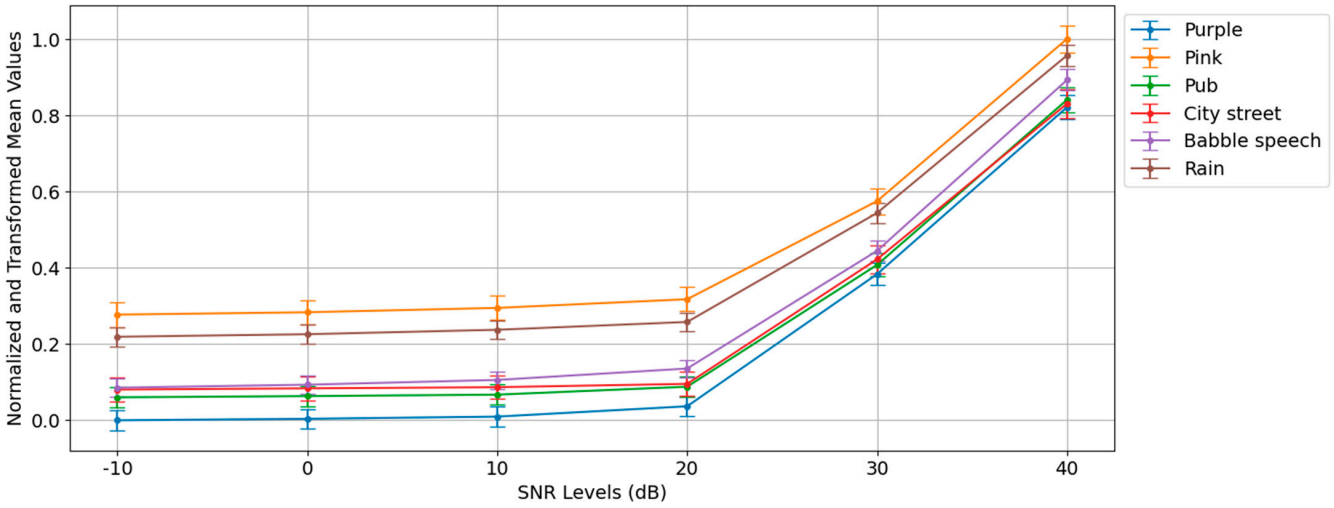


Figure 6. The mean and standard deviation of the normalized and transformed RMSE values for Lombard speech recordings.

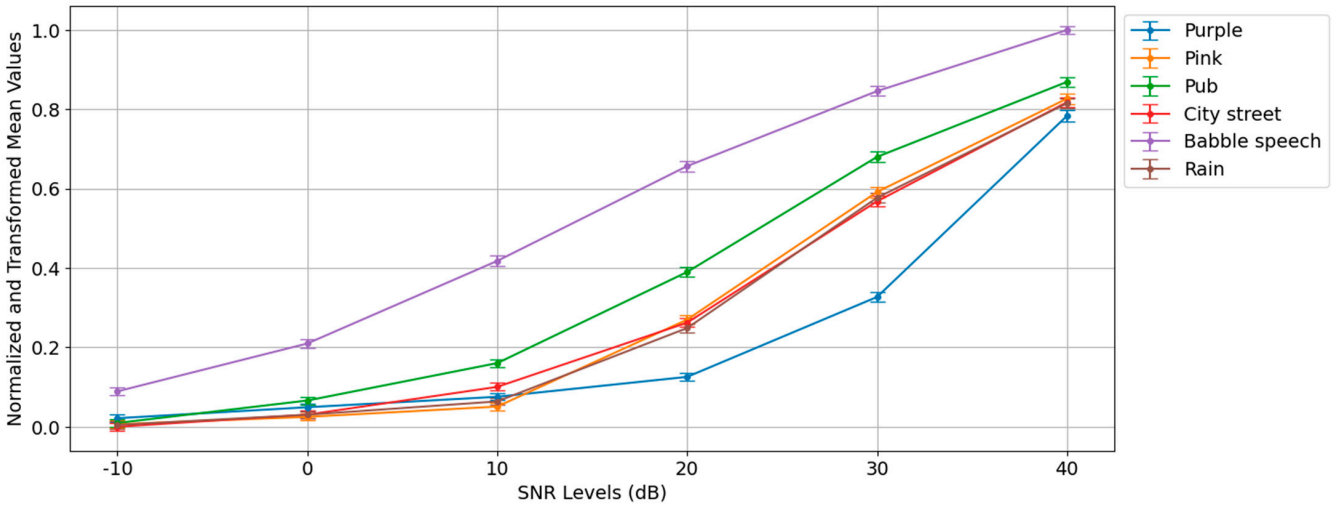


Figure 7. The mean and standard deviation of the normalized and transformed dHash-based values for Lombard speech recordings.

In the case of the RMSE metric, many high standard deviation values were obtained. This indicates that the obtained similarity measures were scattered over a larger distance from the mean and vary widely; therefore, the RMSE measure was not considered further. For the SSIM and dHash-based scores, the best results in terms of less variability were obtained for babble speech noise, followed by recordings mixed with pub noise. Results for city street, pink noise, and rain noise were very similar. Purple noise had a poorer estimate at 20 dB and over, giving a worse result. Further, the spectrum of noise signals was analyzed. The following spectral envelope shape parameters were extracted: Spectral Entropy, Spectral RollOff, and Spectral Brightness. The normalized values are given in Table 4.

Table 4. The normalized spectral characteristics of the noise signals.

Noise Type	Spectral Entropy	Spectral RollOff	Spectral Brightness
Purple noise	1	1	1
Pink noise	0.76	0.29	0.15
Pub noise	0.88	0.51	0.34
City street noise	0.97	0.84	0.84
Babble speech noise	0.82	0.47	0.17
Rain noise	1.00	1.00	1.00

When comparing the spectrum-based values (Table 4) of the noise signal that was analyzed, it was observed that the spectral entropy, which measures spectrum irregularity, reflects the unpredictability of these signals. This may have led to lower SSIM index and dHash-based values for these noises (except for the pink noise). Also, the amount of high-frequency information, which Spectral Brightness and RollOff reflect, directly impacts the SSIM index and dHash-based scores presented in Tables 1 and 3, respectively.

3.2. Modification of the Neutral Speech Samples Based on Noise Profiling

The investigations conducted also included an attempt to modify neutral speech samples based on noise profiling and test them using a speech-quality indicator. Automatic noise profiling, utilizing the developed noise profiling method presented initially in a paper by Korvel et al. [19], and later extended and published by Kąkol et al. [41], was used for this purpose. This method employs machine learning to perform near-real-time noise profiling. The noise recognition model is built upon a Naïve Bayes [42] classifier, using noise signal features derived from the Aurora noise dataset [43]. The target classes used were airport,

babble speech, car noise, exhibition, restaurant, street noise, subway, train, and pink noise. Each recording containing noise was processed as follows:

- For the training process, a signal was divided into a frame length of 2 s to collect the statistical features.
- During each analysis step, the 2-s window was shifted by 0.1 s.

This work aims to modify neutral speech samples based on the noise profiling results obtained. The modifications that were applied to the signal are related to one of the most well-known Lombard speech characteristics—pitch shifting. The F0 is calculated using the following 2nd-degree polynomial equation:

$$y = a_0 + a_1 x + a_2 x^2 \tag{11}$$

where x represents an SNR level, and a_0, a_1, a_2 are coefficients obtained from track change analysis. The polynomial coefficients obtained from track change analysis are given in Table 5. The coefficient of determination (R^2 values) was calculated to indicate the goodness of fit of a curve to the data.

Table 5. Polynomial coefficients and R^2 values obtained from track change analysis.

Type Noise	a_2	a_1	a_0	R^2 Value
Pub noise	−0.0073	0.0113	1.349	0.9865
City street noise	−0.0011	0.0462	1.308	0.987
Babble speech noise	−0.0086	0.0131	1.436	0.9874
Rain noise	−0.0107	0.0016	1.422	0.9869

To compare the results, statistical pitch modification was made, which involved increasing F0 by 10% regardless of the type of noise and its signal-to-noise ratio. F0 trajectory for speech modifications was extracted using the STRAIGHT vocoder. The basic principles of this algorithm can be found in a paper by Kawahara [44]. An overlap-add synthesis using minimum-phase impulse response with group delay manipulations was employed for synthesized speech creation.

A graphical representation of a neutral speech fragment with added nonstationary restaurant noise at 0 dB SNR, displayed together with its synthesized variants using two different F0 modification techniques, i.e., increase F0 by 10% and noise profiling-based F0 modifications, is given in Figure 8.

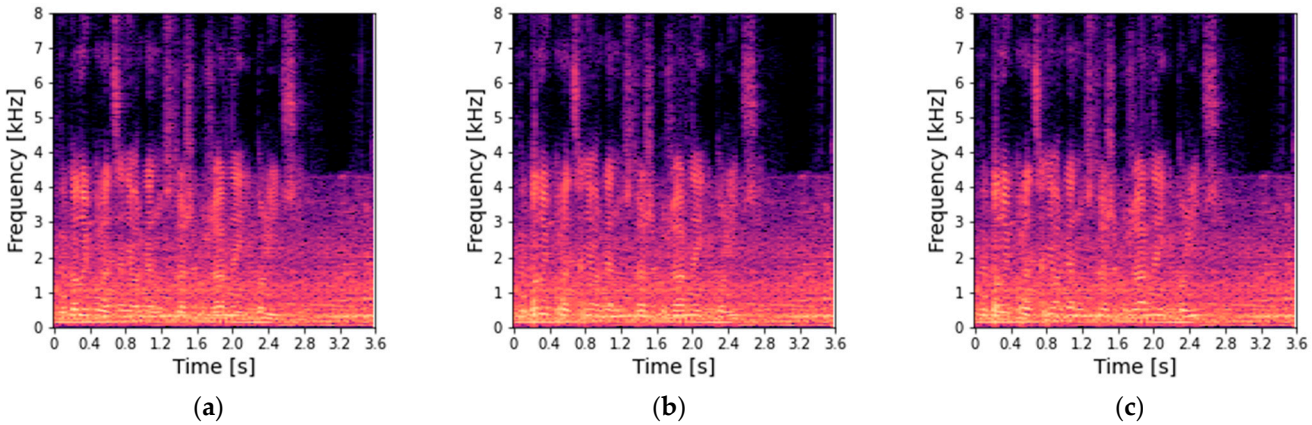


Figure 8. A neutral speech fragment with added nonstationary street noise at 0 dB SNR: (**a**) not modified, (**b**) increase F0 by 10%, (**c**) noise profiling-based F0 modifications. Brighter colors correspond to higher amplitudes, while darker colors correspond to lower amplitudes.

To measure the quality of the synthesized speech signals, the Perceptual Evaluation of Speech Quality (PESQ) test, which was standardized using the ITU-T in Recommendation

P.862 [45,46], was implemented in this research. This implementation enabled the undertaking of PESQ measurements by comparing a degraded speech signal with a reference (undegraded) speech signal. As a result, scores known as Mean Opinion Scores for Listening Quality Objective (MOS-LQO) were generated. These scores represent the perceived quality of the degraded speech.

Estimated MOS-LQO values are given in Table 6, where the highest scores for each noise type and SNR are highlighted in bold font. A graphical representation of the results, along with values of the standard deviations, is shown in Figure 9.

Table 6. The MOS-LQO values (measure produces a score ranging from −0.5 to 4.5, with higher scores indicating better speech quality; such values are highlighted in bold font).

			−10 dB	0 dB	10 dB	20 dB	30 dB	40 dB
Pub noise	Normal speech	Mean	0.925	1.283	2.086	2.824	3.656	4.187
		STD	0.309	0.423	0.227	0.196	0.144	0.070
	Increase F0 by 10%	Mean	0.964	1.351	2.106	2.849	3.680	4.196
		STD	0.343	0.434	0.228	0.193	0.136	0.068
	Noise profiling-based F0 modifications	Mean	**0.969**	**1.479**	**2.162**	**2.899**	**3.703**	**4.196**
		STD	0.358	0.407	0.205	0.182	0.142	0.068
City street noise	Normal speech	Mean	0.878	1.192	1.850	2.589	3.490	4.162
		STD	0.297	0.346	0.245	0.211	0.187	0.083
	Increase F0 by 10%	Mean	0.835	1.173	1.882	2.618	3.521	**4.179**
		STD	0.314	0.373	0.244	0.200	0.181	0.080
	Noise profiling-based F0 modifications	Mean	**0.899**	**1.213**	**1.914**	**2.670**	**3.583**	**4.179**
		STD	0.350	0.326	0.234	0.198	0.164	0.080
Babble speech noise	Normal speech	Mean	**1.278**	1.510	2.308	2.974	3.707	4.175
		STD	0.420	0.458	0.203	0.193	0.119	0.079
	Increase F0 by 10%	Mean	1.271	1.544	2.306	2.992	3.726	**4.186**
		STD	0.352	0.468	0.248	0.191	0.126	0.081
	Noise profiling-based F0 modifications	Mean	1.264	**1.682**	**2.361**	**3.014**	**3.726**	4.175
		STD	0.388	0.449	0.193	0.190	0.126	0.079
Rain noise	Normal speech	Mean	0.747	1.299	2.059	2.861	3.776	4.277
		STD	0.298	0.329	0.232	0.223	0.147	0.066
	Increase F0 by 10%	Mean	0.875	1.339	2.079	2.890	3.805	**4.286**
		STD	0.545	0.295	0.223	0.215	0.136	0.065
	Noise profiling-based F0 modifications	Mean	**0.883**	**1.361**	**2.118**	**2.966**	**3.836**	**4.286**
		STD	0.346	0.281	0.222	0.188	0.129	0.065

Regarding the MOS-LQO values, it was observed that the quality of the synthesized speech in noise was generally enhanced through the utilization of the adopted F0 modifications in the literature, except for city street noise at 0 dB. The noise profiling-based F0 modifications have yielded better scores compared with increasing F0 by 10%. It was also found that across different scenarios, recordings with babble noise generally obtained the highest scores, followed by those mixed with pub noise. This assessment includes all noise conditions considered, except for babble speech at −10 dB (see Table 6 and Figure 9). It should be noted that babble noise produced the fewest frequency changes, as per the track-change analysis. In contrast, this type of noise causes the most problems in the speech-denoising process, as it requires removing one speech source without affecting the other. The method of modification, F0, as proposed in this research, while effective in most scenarios, may require further refinement to handle all types of noise. It should also be researched as to its viability in real-world noise environments.

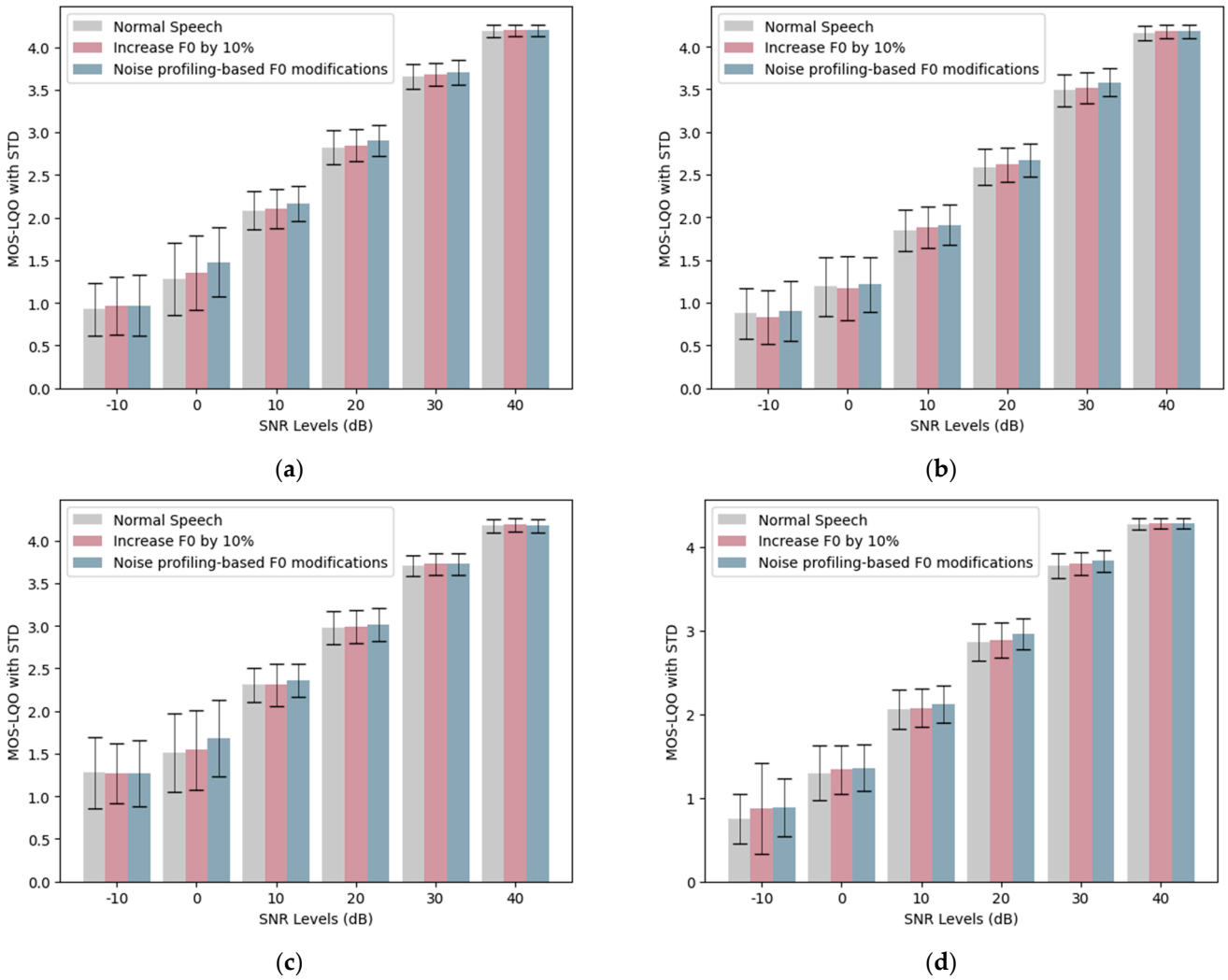


Figure 9. The MOS-LQO values (measure produces a score ranging from −0.5 to 4.5, with higher scores indicating better speech quality): (**a**) Pub noise, (**b**) City street noise, (**c**) Babble speech noise, (**d**) Rain noise.

4. Conclusions

This paper presents the results of a study that examined the spectral characteristics of Lombard speech under noise interference. This study aimed to extend the existing theoretical knowledge on the Lombard effect by analyzing how different types of noise affect it; that is why the experiment was carried out in controlled conditions.

The investigation carried out demonstrated that the highest performance was achieved for babble noise, followed by pub noise. This paper also shows a clear correlation between the SSIM indexes and the spectral characteristics of the noise signal. A relationship between the Spectral Brightness, RollOff, and Entropy of the noise signal and the signal degradation is also observed, i.e., the higher these values are, the more the speech signal is degraded.

The MOS-LQO values indicate that in most cases (except for babble speech noise at 0 dB SNR), the adopted F0 modifications improve the quality of speech synthesized in noise, resulting in the leading position of noise profiling-based F0 modifications. It should, however, be noted that these changes in speech quality are not significant with the F0-applied modifications.

Although Lombard speech processing and synthesis have made many advances in recent years, there is still a need to improve speech synthesis models to be more robust in adverse SNR conditions. Based on the research, further investigation could be done to use

more extensive modifications (not just F0 modifications) to make the model more robust in adverse noise conditions. Our previous study revealed that the analysis conducted separately for individual speakers shows remarkable differences between them [47]. In the future, the testing of the idea of a male–female division is highlighted as an important thread to be checked [48]. Also, deep-learning algorithms instead of the state-of-the-art vocoder can be applied to such a task. This way, a more extensive repository of records may be obtained; however, the results of the synthesized LE, incorporated into natural speech, should be checked in subjective tests.

Moreover, the paper does not consider the case of mixed noises, which may occur in real life. Therefore, this is another issue that requires further investigation. Also, future analysis could be conducted to investigate the impact of high-frequency information on speech intelligibility in noisy environments. This will include evaluating the benefits of higher sampling rates and the presence of high-frequency noise components, which may more accurately reflect real-world listening scenarios, such as trying to understand speech in a noisy environment.

There are some general outcomes of the study performed:

1. Implications for Speech Synthesis and Recognition Technologies:
   - The findings from this study could impact the development of more advanced speech synthesis and recognition systems. By incorporating the principles of the Lombard effect, these systems could adapt to noisy environments, enhancing their effectiveness in real-world scenarios like crowded public spaces or vehicles.
2. Application in Assistive Technologies:
   - This research holds potential benefits for assistive communication technologies, especially for individuals with hearing impairments. The enhanced clarity and intelligibility of speech generated through LE-influenced methods could improve communication in challenging auditory environments.
3. Future Research Directions:
   - There is scope for exploring the application of the Lombard effect in different languages, as speech characteristics and responses to noise may vary across languages.
   - Further research could also delve into the subjective perception of LE-modified speech by listeners, particularly in terms of naturalness and ease of understanding.
4. Integration with Ambient Intelligence Systems:
   - This study's methodology and findings can be integrated into ambient intelligence systems for smarter environment-responsive communication aids, for instance, in smart homes or workplaces, where the system may automatically adjust communication modalities based on the detected noise levels.
5. Challenges and Limitations:
   - This study highlights the challenges in accurately mimicking the natural Lombard effect, particularly in maintaining the naturalness of speech while enhancing intelligibility.
   - Limitations in current speech synthesis technologies in replicating complex human vocal nuances in varying noise conditions need addressing.
6. Broader Societal Impact:
   - This research can contribute to enhancing public safety communications, like emergency announcements in noisy environments, by ensuring clearer and more intelligible speech transmission.
   - It also has implications for improving communication in educational settings, especially in noisy classrooms or during outdoor activities.

By exploring these additional aspects, the conclusions of the paper can be broadened to encompass a wider range of implications, applications, and future research directions.

Author Contributions: Conceptualization, G.K. and B.K.; methodology, G.K. and K.K.; software, G.K.; validation, P.T. and B.K.; investigation, G.K. and K.K.; data curation, G.K. and B.K.; figure preparation, G.K. and K.K.; writing—original draft preparation, G.K.; writing—review and editing, P.T. and B.K.; supervision, P.T. and B.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Social Fund under the No. 09.3.3-LMT-K-712 "Development of Competences of Scientists, other Researchers and Students through Practical Research Activities" measure.

Data Availability Statement: Data are contained within the article.

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

Appendix A

Algorithm A1: The pseudo-code of the peak matching.

INPUT:
ω_n^l—the frequency on frame l
ω_m^{l+1}—the frequency on frame $l+1$
N—the total number of peaks in frame l
M—the total number of peaks in frame $l+1$
$n = 0, \ldots, N-1$
$m = 0, \ldots, M-1$
$p = 0, \ldots, M-1$
$p \neq m$
for each frequency in frame l **do**
STEP 1. **if** $\left|\omega_n^l - \omega_m^{l+1}\right| \geq \Delta$ **then**
ω_n^l is matched to itself in frame $l+1$
the amplitude of ω_n^l is set to zero
else
if $\left(\left|\omega_n^l - \omega_m^{l+1}\right| < \left|\omega_n^l - \omega_p^{l+1}\right| < \Delta\right)$ **then**
$\omega_m^{l+1}$is declared to be a candidate to ω_n^l
end if
end if
STEP 2. **if** $\left(\left|\omega_m^{l+1} - \omega_n^l\right| < \left|\omega_m^{l+1} - \omega_{p+1}^l\right|$, where $p > i$) **then**
ω_n^l is matched to ω_m^{l+1}
else
if ω_{m-1}^{l+1} exists **then**
if $\left|\omega_n^l - \omega_{m-1}^{l+1}\right| < \Delta$ then
ω_n^l is matched to ω_{m-1}^{l+1}
else
ω_n^l is matched to itself in frame $l+1$
the amplitude of ω_n^l is set to zero
end if
end if
STEP 3. **for** the remaining frequencies in frame $l+1$
frequencies are created in frame l with zero amplitude
the match is made

The comments on the algorithm:

✓ If the frequencies are matched, they are eliminated from further consideration.
✓ Δ denotes a matching interval [21]

References

1. Lombard, E. Le signe de l'elevation de la voix. *Ann. Mal. L'Oreille Larynx* **1911**, *37*, 101–119.
2. Marxer, R.; Barker, J.; Alghamdi, N.; Maddock, S. The impact of the Lombard effect on audio and visual speech recognition systems. *Speech Commun.* **2018**, *100*, 58–68. [CrossRef]
3. Available online: https://en.wikipedia.org/wiki/Lombard_effect (accessed on 3 November 2023).
4. Zollinger, S.A.; Brumm, H. The lombard effect. *Curr. Biol.* **2011**, *21*, 614–615. [CrossRef] [PubMed]
5. Available online: https://www.rockfon.co.uk/about-us/blog/2023/lombard-effect-solutions/ (accessed on 3 November 2023).
6. Available online: https://www.fohlio.com/blog/psychology-restaurant-interior-design-part-4-restaurant-acoustics (accessed on 3 November 2023).
7. Hansen, J.H.; Lee, J.; Ali, H.; Saba, J.N. A speech perturbation strategy based on "Lombard effect" for enhanced intelligibility for cochlear implant listeners. *J. Acoust. Soc.Am.* **2020**, *147*, 1418–1428. [CrossRef] [PubMed]
8. Vlaj, D.; Kacic, Z. The influence of Lombard effect on speech recognition. *Speech Technol.* **2011**, 1998–2001.
9. Kang, T.; Dinh, A.D.; Wang, B.; Du, T.; Chen, Y.; Chau, K. Optimization of a Real-Time Wavelet-Based Algorithm for Improving Speech Intelligibility. *arXiv* **2022**, arXiv:2202.02545.
10. Bollepalli, B.; Juvela, L.; Airaksinen, M.; Valentini-Botinhao, C.; Alku, P. Normal-to-lombard adaptation of speech synthesis using long short-term memory recurrent neural networks. *Speech Commun.* **2019**, *110*, 64–75. [CrossRef]
11. Suni, A.; Karhila, R.; Raitio, T.; Kurimo, M.; Vainio, M.; Alku, P. Lombard modified text-to-speech synthesis for improved intelligibility: Submission for the hurricane challenge 2013. In Proceedings of the Interspeech, Lyon, France, 25–29 August 2013; pp. 3562–3566. [CrossRef]
12. Uma Maheswari, S.; Shahina, A.; Nayeemulla Khan, A. Understanding lombard speech: A review of compensation techniques towards improving speech based recognition systems. *Artif. Intell. Rev.* **2021**, *54*, 2495–2523. [CrossRef]
13. Li, G.; Hu, R.; Zhang, R.; Wang, X. A mapping model of spectral tilt in normal-to-lombard speech conversion for intelligibility enhancement. *Multimed. Tools Appl.* **2020**, *79*, 19471–19491. [CrossRef]
14. Kakol, K.; Korvel, G.; Kostek, B. Improving objective speech quality indicators in noise conditions. In *Data Science: New Issues, Challenges and Applications*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 199–218.
15. Bollepalli, B.; Juvela, L.; Alku, P. Lombard Speech Synthesis Using Transfer Learning in a Tacotron Text-to-Speech System. In Proceedings of the Interspeech, Graz, Austria, 15–19 September 2019; pp. 2833–2837. [CrossRef]
16. Hu, Q.; Bleisch, T.; Petkov, P.; Raitio, T.; Marchi, E.; Lakshminarasimhan, V. Whispered and Lombard Neural Speech Synthesis. In Proceedings of the 2021 IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021; pp. 454–461. [CrossRef]
17. Paul, D.; Shifas, M.P.; Pantazis, Y.; Stylianou, Y. Enhancing speech intelligibility in text-to-speech synthesis using speaking style conversion. *arXiv* **2020**, arXiv:2008.05809.
18. Korvel, G.; Kąkol, K.; Kurasova, O.; Kostek, B. Evaluation of Lombard speech models in the context of speech in noise enhancement. *IEEE Access* **2020**, *8*, 155156–155170. [CrossRef]
19. Korvel, G.; Kąkol, K.; Treigys, P.; Kostek, B. Investigating Noise Interference on Speech Towards Applying the Lombard Effect Automatically. In Proceedings of the Foundations of Intelligent Systems: 26th International Symposium, ISMIS 2022, Cosenza, Italy, 3–5 October 2022; Springer International Publishing: Cham, Switzerlands, 2022; pp. 399–407.
20. Novitasari, S.; Sakti, S.; Nakamura, S. Dynamically adaptive machine speech chain inference for tts in noisy environment: Listen and speak louder. In Proceedings of the Interspeech, Brno, Czechia, 30 August–3 September 2021; pp. 4124–4128.
21. Yue, F.; Deng, Y.; He, L.; Ko, T.; Zhang, Y. Exploring machine speech chain for domain adaptation. In Proceedings of the ICASSP 2022-IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 6757–6761.
22. Chavdar, M.; Kartalov, T.; Ivanovski, Z.; Taskovski, D.; Gerazov, B. SCarrie: A Real-Time System for Sound Event Detection for Assisted Living. In Proceedings of the 30th International Conference on Systems, Signals and Image Processing (IWSSIP), Ohrid, North Macedonia, 27–29 June 2023; pp. 1–5. [CrossRef]
23. McAulay, R.; Quatieri, T. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans.Acoust. Speech Signal Process.* **1986**, *34*, 744–754. [CrossRef]
24. Lampert, T.A.; O'Keefe, S.E. On the detection of tracks in spectrogram images. *Pattern Recognit.* **2013**, *46*, 1396–1408. [CrossRef]
25. Bhattacharjee, M.; Prasanna, S.M.; Guha, P. Speech/music classification using features from spectral peaks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1549–1559. [CrossRef]
26. Baese-Berk, M.M.; Levi, S.V.; Van Engen, K.J. Intelligibility as a measure of speech perception: Current approaches, challenges, and recommendations. *J. Acoust. Soc. Am.* **2023**, *153*, 68–76. [CrossRef]
27. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Tions Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
28. Peng, J.; Shi, C.; Laugeman, E.; Hu, W.; Zhang, Z.; Mutic, S.; Cai, B. Implementation of the structural similarity (ssim) index as a quantitative evaluation tool for dose distribution error detection. *Med. Phys.* **2020**, *47*, 1907–1919. [CrossRef] [PubMed]
29. Zini, S.; Bianco, S.; Schettini, R. Deep residual autoencoder for blind universal jpeg restoration. *IEEE Access* **2020**, *8*, 63283–63294. [CrossRef]
30. Le, H.; Samaras, D. Shadow removal via shadow image decomposition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8578–8587.

31. Shao, Z.; Cai, J.; Fu, P.; Hu, L.; Liu, T. Deep learning-based fusion of Landsat-8 and Sentinel-2 images for a harmonized surface reflectance product. *Remote Sens. Environ.* **2019**, *235*, 111425. [CrossRef]
32. Available online: https://www.okta.com/identity-101/hashing-algorithms (accessed on 3 November 2023).
33. Available online: https://builtin.com/cybersecurity/what-is-hashing (accessed on 3 November 2023).
34. Available online: https://cheapsslsecurity.com/blog/decoded-examples-of-how-hashing-algorithms-work (accessed on 3 November 2023).
35. Xue, M.; He, C.; Wang, J.; Liu, W. Backdoors hidden in facial features: A novel invisible backdoor attack against face recognition systems. *Peer-to-Peer Netw. Appl.* **2021**, *14*, 1458–1474. [CrossRef]
36. Song, W.; Gao, Z.; Dian, R.; Ghamisi, P.; Zhang, Y.; Benediktsson, J.A. Asymmetric hash code learning for remote sensing image retrieval. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
37. Chen, Y.; Tang, Y.; Huang, J.; Xiong, S. Multi-scale Triplet Hashing for Medical Image Retrieval. *Comput. Biol. Med.* **2023**, *155*, 106633. [CrossRef] [PubMed]
38. Yang, X.; Feng, L.; Lu, T.; Dong, Q. Application of image hash algorithm in copyright protection system. In Proceedings of the Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT), Harbin, China, 3–5 December 2021; Volume 12167, pp. 800–807.
39. Joshi, A.; Shet, A.V.; Thambi, A.S.; Sunitha, R. Quality Improvement of Image Datasets using Hashing Techniques. In Proceedings of the 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bengaluru, India, 27–28 January 2023; pp. 18–23.
40. Buchner, J. A Python Perceptual Image Hashing Module: Imagehash. 2020. Available online: https://github.com/JohannesBuchner/imagehash (accessed on 3 November 2023).
41. Kąkol, K.; Korvel, G.; Kostek, B. Noise profiling for speech enhancement employing machine learning models. *J. Acoust. Soc. Am.* **2022**, *152*, 3595–3605. [CrossRef] [PubMed]
42. Barber, D. *Bayesian Reasoning and Machine Learning*; Cambridge University Press: Cambridge, UK, 2012; ISBN 978-0-521-51814-7.
43. Hirsch, H.G.; Pearce, D. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In Proceedings of the ASR2000-Automatic Speech Recognition: Challenges for the New Millenium ISCA Tutorial and Research Workshop (ITRW), Paris, France, 18–20 September 2000; pp. 181–188.
44. Kawahara, H. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoust. Sci. Technol.* **2006**, *27*, 349–353. [CrossRef]
45. Union, I.T. *Wideband Extension to Recommendation p. 862 for the Assessment of Wideband Telephone Networks and Speech Codecs*; International Telecommunication Union: Geneva, Switzerland, 2007.
46. Beerends, J.G.; Hekstra, A.P.; Rix, A.W.; Hollier, M.P. PESQ, the new ITU standard for objective measurement of perceived speech quality—Part II: Perceptual model. *J. Audio Eng. Soc.* **2002**, *50*, 765–778.
47. Piotrowska, M.; Korvel, G.; Kostek, B.; Ciszewski, T.; Czyżewski, A. Machine learning-based analysis of English lateral allophones. *Int. J. Appl. Math. Comput. Sci.* **2019**, *29*, 393–405. [CrossRef]
48. Alghamdi, N.; Maddock, S.; Marxer, R.; Barker, J.; Brown, G.J. A corpus of audio-visual Lombard speech with frontal and profile views. *J. Acoust. Soc. Am.* **2018**, *143*, EL523–EL529. [CrossRef]