Accepted manuscript of: Korvel, G., Kąkol, K., Treigys, P., Kostek, B. (2022). Investigating Noise Interference on Speech Towards Applying the Lombard Effect Automatically. In: Ceci, M., Flesca, S., Masciari, E., Manco, G., Raś, Z.W. (eds) Foundations of Intelligent Systems. ISMIS 2022. Lecture Notes in Computer Science(), vol 13515. Springer, Cham. https://doi.org/10.1007/978-3-031-16564-1_38

Investigating noise interference on speech towards applying the Lombard effect automatically

Gražina Korvel[1], Krzysztof Kąkol[2], Povilas Treigys[1], Bożena Kostek[3]
[1]Institute of Data Science and Digital Technologies, Vilnius University, Vilnius, Lithuania
[2]PGS Software, Wrocław, Poland
[3]Audio Acoustics Laboratory, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdansk, Poland
grazina.korvel@mif.vu.lt

Abstract. The aim of this study is two-fold. First, we perform a series of experiments to examine the interference of different noises on speech processing. For that purpose, we concentrate on the Lombard effect, an involuntary tendency to raise speech level in the presence of background noise. Then, we apply this knowledge to detecting speech with the Lombard effect. This is for preparing a dataset for training a machine learning-based system for automatic speech conversion, mimicking a human way to make speech more intelligible in the presence of noise, i.e., to create Lombard speech. Several spectral descriptors are analyzed in the context of Lombard speech and various types of noise. In conclusion, pub-like and babble noises are most similar when comparing Spectral Entropy, Spectral RollOff, and Spectral Brightness. The larger values of these spectral descriptors, the more the speech-in-noise signal is degraded. To quantify the effect of noise on speech, containing the Lombard effect, an average formant track error is calculated as an objective image quality metric. For image quality assessment Structural SIMilarity (SSIM) index is employed.

Keywords: Lombard effect, noise background, Structural SIMilarity (SSIM) index.

1 Introduction

A number of approaches to robust speech processing are seen in the literature and practical solutions. Despite this, when we refer 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. This study builds on the idea of incorporating the Lombard effect (LE) into speech in adverse environments. The Lombard phenomenon, named after the French otolaryngologist Étienne Lombard, occurs in speech production in the presence of noise [1]. He observed that when patients were exposed to loud noise during a conversation, they involuntary raised their voice level and speech became more intelligible. So, to build a human-centric system with ambient intelligence to generate speech with LE for better intelligibility, first, we need to learn about noise interference on speech characteristics. Second, to enable the system

to generate Lombard speech when noise is detected and correctly labeled, the interference sound recognition model should be trained on speech with this phenomenon present in it. Moreover, it is evident that by applying a deep model, a large amount of data with the Lombard effect is needed.

Since the discovery of LE, this phenomenon has been extensively studied by a wide range of specialists to find solutions to improve the performance of automatic speech recognition systems in noisy environments [2] or increase speech intelligibility by converting the speaking style from normal to Lombard speech [3, 4]. The idea is that LE may be applied to speech synthesizers, allowing them to adapt to noisy conditions [5, 6, 7]. It should be noted that text-to-speech systems adapt to the noise condition during the training process. In contrast, noise profiling still needs to be examined, though some research has already been carried out in this direction with promising results [8, 9].

As already mentioned, our long-term goal is to build a human-centric interface for ambient intelligence to generate speech with the Lombard effect, which could perform automatic adaptation during noise inference. This research investigates the effect of noise interference on Lombard speech. We need such analysis to determine whether speech available on the Internet is with LE or not because we want to use them for the deep network training. So, we investigate to what extent and how the clean speech with LE differs from Lombard speech in noise. For this purpose, specifically, rapidly changing areas of speech such as voiced/unvoiced transitions are examined. To indicate such changes in spectral energy, frequency tracks should be estimated. Both the location and the number of peaks are important in this context. The study deals with various additive noises and different SNR (Signal-to-Noise) levels.

2 Estimation of frequency tracks

We conduct the speech analysis that is based on the signal intensity at each time-frequency point. The process of determining frequency tracks in a speech signal is shifted to finding them in a spectrogram, a visual representation of the distribution of signal acoustic energy across frequencies and over time. The darkness of the energy bands is used to estimate the signal intensity. The spectrogram creation process consists of the calculation of the discrete Fourier transform of each short-time frame of speech signal:

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

where $X_l(k)$ are Fourier transform coefficients ($k = 0, \dots, N_{FT} - 1, N_{FT}$ is the number of Fourier transform coefficients), $x_l(n)$ – the samples of lth short-time frame of signal ($l = 0, \dots, L - 1$, and L denotes the number of short-time frames), N – the length of the signal, $w(n) = 0{,}54 - 0.46\mathrm{con}(2\pi n/N - 1)$ is the Hamming window function, and j is the imaginary unit.

The obtained values are then collected together, and a spectrogram image is built up. A graphical representation of the spectrogram obtained is given in Fig. 1, where both the clean Lombard speech fragment and the same speech fragment corrupted by nonstationary street noise at 0 dB SNR are displayed. For the purpose of this analysis, the spectrogram representation is generated using Hamming windows of size 512. This

window size gives smoothed Fourier spectrum. At the same time, the frequency resolution is sufficient for frequency tracking. An overlap of 256 is used to avoid losing part of the information due to the window operation.

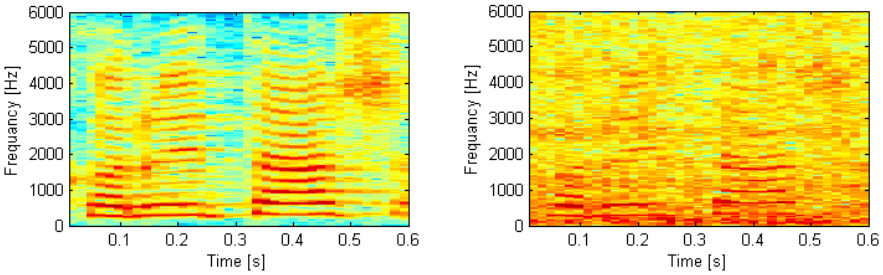

Fig. 1. The spectrogram of clean (the left side) and the noisy (the right side) speech signal.

Various tracking methods and their modifications were proposed [10, 11]. In this research, we used a classical algorithm proposed by McAulay and Quartieri (McA-Q) [12]. The detection of frequency tracks is performed in spectrograms. First, all local maxima of the spectrogram are detected in each short-time frame l. These maxima are called peaks. The estimated peaks, i.e., the amplitudes and their frequencies, are then passed to the tracking algorithm, whose aim is to remove partial trajectories. According to the McA-Q algorithm, frame-to-frame peak matching is performed. The process of matching each spectrum peak in frame l to the peaks in frame $l + 1$, is presented by the following pseudo-code, shown in the algorithmic form.

The result of applying the tracker to the Lombard speech signal is shown in Fig. 2.

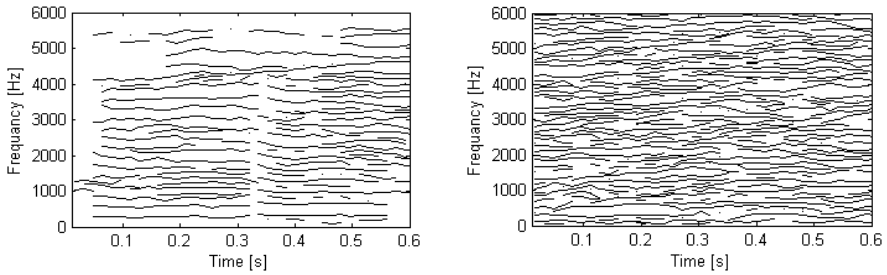

Fig. 2. The estimated frequency tracks of clean (the left side of the figure) and the noisy (the right side) Lombard speech signal.

As we see from the pseudo-code given above, matching 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 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.

Algorithm

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** $|\omega_n^l - \omega_m^{l+1}| \geq \Delta$ **then**
ω_n^l is matched to itself in a frame $l+1$
the amplitude of ω_n^l is set to zero
else
if $(|\omega_n^l - \omega_m^{l+1}| < |\omega_n^l - \omega_p^{l+1}| < \Delta)$ **then**
ω_m^{l+1} is declared to be a candidate to ω_n^l
end if
end if
STEP 2. **if** $(|\omega_m^{l+1} - \omega_n^l| < |\omega_m^{l+1} - \omega_{p+1}^l|$, where $p > i)$ **then**
ω_n^l is matched to ω_m^{l+1}
else
if ω_{m-1}^{l+1} exists **then**
if $|\omega_n^l - \omega_{m-1}^{l+1}| < \Delta$ then
ω_n^l is matched to ω_{m-1}^{l+1}
else
ω_n^l is matched to itself in a 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 [12]

Fig. 2 shows the frequency tracks of the clean speech segment and the same speech segment corrupted by nonstationary street noise at 0 dB SNR. A ratio of 0 dB indicates the signal level is the same as the noise level; therefore, degradation of formant tracks of noisy speech is visible.

3 Image comparison technique

To quantify the effect of noise on speech, containing the Lombard effect, an average formant track error is calculated as an objective image quality metric. For image quality assessment Structural SIMilarity (SSIM) index is calculated. The SSIM index was developed by Wang et al. [13] to evaluate the quality of two images based on the perspective of image formation, i.e., the image luminance, contrast, and structural similarity. The above-mentioned advantages of this method make it sensitive to changes in the image, which is very important in our study. It should also be noted that SSIM is widely used as the quality indicator of the images being compared [14, 15].

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two nonnegative image signals. The structural SSIM index is calculated by the following formula [13]:

$$S(\boldsymbol{x},\boldsymbol{y}) = [l(\boldsymbol{x},\boldsymbol{y})]^{\alpha} \cdot [c(\boldsymbol{x},\boldsymbol{y})]^{\beta} \cdot [s(\boldsymbol{x},\boldsymbol{y})]^{\gamma} \tag{2}$$

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

$$l(\boldsymbol{x},\boldsymbol{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{3}$$

$$c(\boldsymbol{x},\boldsymbol{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{4}$$

$$s(\boldsymbol{x},\boldsymbol{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{5}$$

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 [13]. The overall similarity measure SSIM is in the range -1 to 1. A value of 1 indicates an ideal agreement between two images, while a value of -1 indicates the given images are very different. In this research, using the SSIM index, we calculated the difference between the image of estimated frequency tracks of the clean speech signal and that of the noisy speech signal.

4 Experimental setup

Eight speakers (four males and four females) were separately asked to utter fifteen sentences. The speakers were untrained healthy native students of the Gdansk University of Technology. Each speaker was asked to repeat each sentence twice under a different acoustic treatment (in a room with and without an acoustically treated interior that suppresses reverberation). To obtain the Lombard effect while speaking, closed headphones played back the interfering noise were used. The recordings were split into smaller segments, the length of which was 1 second. As a result, 2719 recordings of the acoustically treated room and 3109 ones of the room without acoustic treatment were used in the experiment. This is to balance stratification sampling. Moreover, in this

research, we employ four real-life noise recordings, including babble speech (i.e., a mix of many talkers), city streets, rain, and pub that were added to audio data. These recordings were taken from the YouTube platform. The sampling rate of speech and noise signal has been adjusted to 16 kHz before the test.

5 Experimental results

The experiment is designed to measure the influence of noise interference on the frequency tracks of Lombard speech. 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). The investigation carried out concerned both acoustically treated and untreated rooms, however, the results were quite similar. As the untreated room conditions are closer to a typical real-life scenario, therefore these results are shown in Table 1. The SSIM index values indicating the correspondence between the shape of a speech signal with LE and its noisy version on different SNR conditions are contained in Table 1. The graphical representation of the results obtained is given in Fig. 3.

The best results were obtained for babble speech noise, followed by recordings mixed with pub noise. Results for city street and rain noises are very similar. For city street noise, there is a slightly higher estimate at the 10 dB condition towards a better result. Further, we analyzed the spectrum of noise signals. The following spectral envelope shape parameters were extracted: Spectral Entropy, Spectral RollOff, and Spectral Brightness. The normalized values are given in Table 2.

When comparing the spectrum-based values (see Table 2) of the noise signal analyzed, we can observe that the spectral entropy, which gives a measure of spectrum irregularity [16], reflects the unpredictability of these signals. This may have led to lower values of the SSIM index values for these noises. Also, the amount of high-frequency information, which is reflected by Spectral Brightness and RollOff, has a direct impact on the SSIM index presented in Table 3.

Table 1. The SSIM index values for recordings (a room without acoustic treatment)

		-10 dB	0 dB	10 dB	20 dB	30 dB	40 dB
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
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

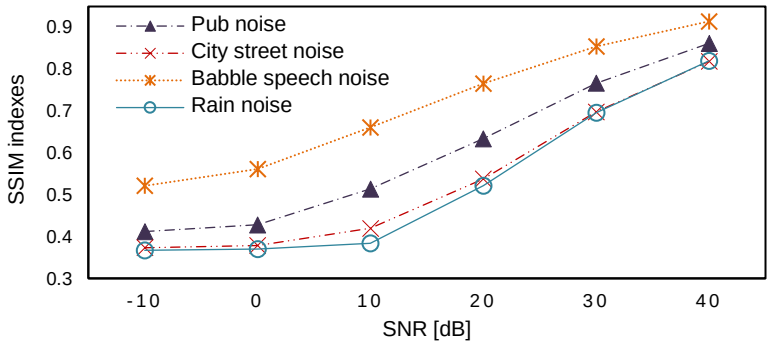

Fig. 3. The SSIM index values for recordings of a room without acoustic treatment.

Table 2. The normalized spectral characteristics of the noise signals

	Spectral Entropy	Spectral RollOff	Spectral Brightness
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

The investigations carried out also contain the first attempt to automatic noise profile based on recordings contained in the MODALITY multimodal corpus of English speech recordings [17]. Based on the frequency characteristics, the classification model was built. For that purpose, Naïve Bayes [18] algorithm was employed. The following target classes were used: airport, babble speech, car noise, exhibition, restaurant, street noise, subway, train, and pink noise. The test was performed only for a 2-seconds frame, and the window was moved by 2 seconds. An example of when the recording was classified as "street," which is the correct classification, is given in Fig. 4.

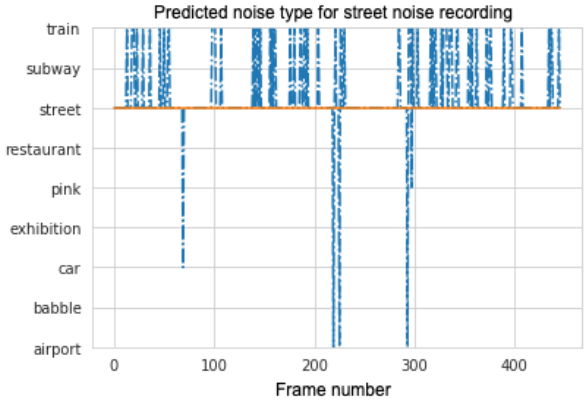

Fig. 4. Classification results on the real-world recordings − the solid line represents the classification in the averaging mode, while the dashed line represents the momentary classification.

In the context of noise profiling, the model's usefulness is measured by evaluating its stability, not the correctness of classification. It can be seen that this process of classification fluctuates while the averaging mode is stable (dashed and solid lines in Fig.4).

6 Conclusions

By analyzing the impact of the noise interference on the Lombard effect, we believe that the existing theoretical background is extended. In this study, we have pointed out that building a human-centric interface for ambient intelligence is an extension of speech processing. The paper shows the outcome of the study that analyzes Lombard speech to understand how the spectral characteristics are affected by noise interference.

The analysis presented in this paper shows that the best results are obtained for babble speech noise, followed by recordings mixed with pub noise. When comparing the spectrum-based values of the noise signal analyzed, there is a clear correlation between the obtained SMM indexes and the obtained spectral characteristics. The greater the Spectral Brightness, RollOff, and Entropy of the interference noise signal, the more the speech signal is degraded.

The influence of noise interference was tested on Lombard speech through an experiment in an acoustically treated room. In real-life, different kinds of noise can be intermingled. The model was not tested against such a combination of noises. The first attempt to profiling noise automatically revealed that LE is applicable in this case. However, this issue needs to be further investigated, which we intend to do in the future.

It is envisioned that the results of this analysis allow for developing a method of monitoring and enhancing speech automatically in the presence of noise. The ultimate goal is to prepare a system capable of synthetically generating Lombard speech through noise profiling.

Acknowledgments

This research is 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.

References

1. Lombard, E.: Le signe de l'elevation de la voix. Ann. Mal. de L'Oreille et du Larynx, 101–119 (1911)Zollinger, S.A., Brumm, H.: The lombard effect. Current Biology 21(16), 614–615 (2011).
2. Uma Maheswari, S., Shahina, A., Nayeemulla Khan, A.: Understanding lombard speech: a review of compensation techniques towards improving speech based recognition systems. Artificial Intelligence Review 54(4), 2495–2523 (2021).

3. Li, G., Hu, R., Zhang, R., Wang, X.: A mapping model of spectral tilt in normal-to-lombard speech conversion for intelligibility enhancement. Multimedia Tools and Applications 79(27), 19471–19491 (2020).
4. Kakol, K., Korvel, G., Kostek, B.: Improving objective speech quality indicators in noise conditions. In: Data Science: New Issues, Challenges and Applications, 199–218. Springer (2020).
5. 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 Communication 110, 64–75 (2019).
6. Paul, D., Shifas, M.P., Pantazis, Y., Stylianou, Y.: Enhancing speech intelligibility in text-to-speech synthesis using speaking style conversion. arXiv preprint arXiv:2008.05809 (2020).
7. Korvel, G., Kąkol, K., Kurasova, O., Kostek, B.: Evaluation of Lombard speech models in the context of speech in noise enhancement, IEEE access, 8, 155156-155170 (2020).
8. Novitasari, S., Sakti, S., Nakamura, S.: Dynamically adaptive machine speech chain inference for tts in noisy environment: Listen and speak louder. Proc. Interspeech 2021, 4124–4128 (2021).
9. Yue, F., Deng, Y., He, L., Ko, T., Zhang, Y.: Exploring machine speech chain for domain adaptation. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6757–6761 (2022).
10. Lampert, T.A., O'Keefe, S.E.: On the detection of tracks in spectrogram images. Pattern recognition 46(5), 1396–1408 (2013).
11. Bhattacharjee, M., Prasanna, S.M., Guha, P.: Speech/music classification using features from spectral peaks. IEEE/ACM Transactions on Audio, Speech, and Language Processing 28, 1549–1559 (2020).
12. McAulay, R., Quatieri, T.: Speech analysis/synthesis based on a sinusoidal representation. IEEE Transactions on Acoustics, Speech, and Signal Processing 34(4), 744–754 (1986).
13. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing 13(4), 600–612 (2004).
14. 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. Medical physics 47(4), 1907–1919 (2020).
15. Zini, S., Bianco, S., Schettini, R.: Deep residual autoencoder for blind universal jpeg restoration. IEEE Access 8, 63283–63294 (2020).
16. Wei, Y., Zeng, Y., Li, C.: Single-channel speech enhancement based on subband spectral entropy. Journal of the Audio Engineering Society 66(3), 100–113 (2018).
17. Czyzewski, A., Kostek, B., Bratoszewski, P., Kotus, J., Szykulski, M.: An audio-visual corpus for multimodal automatic speech recognition, Journal of Intelligent Information Systems, Vol. 49, No. 2, 1–26 (2017).
18. Barber, D.: Bayesian Reasoning and Machine Learning. Cambridge University Press. ISBN 978-0-521-51814-7 (2012).