IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 29, 2020, accepted August 4, 2020, date of publication August 10, 2020, date of current version September 3, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.3015421*

# Evaluation of Lombard Speech Models in the Context of Speech in Noise Enhancement

**GRAŽINA KORVEL[1], KRZYSZTOF KĄKOL[2,3], OLGA KURASOVA[1], AND BOŻENA KOSTEK[2], (Senior Member, IEEE)**

[1]Institute of Data Science and Digital Technologies, Vilnius University, 08412 Vilnius, Lithuania
[2]Audio Acoustics Laboratory, Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, 80-233 Gdańsk, Poland
[3]PGS Software S.A., 50-086 Gdańsk, Poland

Corresponding author: Bożena Kostek (bokostek@audioacoustics.org)

This work was supported in part by the European Social Fund (Development of Competences of Scientists, other Researchers and Students through the Practical Research Activities Measure) under Grant 09.3.3-LMT-K-712, and in part by the Statutory Funds of Electronics, Telecommunications and Informatics Faculty, Gdańsk University of Technology.

**ABSTRACT** The Lombard effect is one of the most well-known effects of noise on speech production. Speech with the Lombard effect is more easily recognizable in noisy environments than normal natural speech. Our previous investigations showed that speech synthesis models might retain Lombard-effect characteristics. In this study, we investigate several speech models, such as harmonic, source-filter, and sinusoidal, applied to Lombard speech in the context of speech enhancement. For this purpose, 100 utterances of natural speech, and 100 with the Lombard effect induced are used. The goal of this study is to check to what extent speech utterances based on these models are recognizable and at what SNR (Signal-to-Noise Ratio) level threshold a particular model stops working. For this purpose, the synthesized models and Lombard speech are mixed with babble speech and street noise recordings with different SNRs. The quality of these models is measured, employing objective indicators as well as subjective tests. Since there is no standardized measure to apply to enhanced speech, an objective measure of assessing the speech quality of a model synthesizing Lombard speech characteristics, based on a feature vector, is proposed. Our approach is then compared with the standardized metric used in telecommunications as well as with subjective test results. The experimental investigations show the superiority of the source-filter models applied to synthesize Lombard speech over other models utilized. Also, the measure proposed correlates more closely with the results of the subjective evaluation than the outcomes from the ITU-T P.563 recommendation. This was checked with a ANOVA statistical analysis.

**INDEX TERMS** Lombard speech, quality of experience, speech modeling techniques.

## I. INTRODUCTION

When it comes to speech quality evaluation prediction, typically, two techniques are employed: objective measures and subjective test results. The International Telecommunication Union (ITU) brings several methods for speech quality evaluation. In some areas of telecommunications, objective measures are well-established [1], and they are compared against subjective tests, which are also standardized (e.g., MUSHRA test [2]). However, in some applications, there are no existing objective measures that are well-adapted to the given field [3]. The area of Lombard speech is one such an example [4]. Therefore, the primary focus of this study is to propose an objective measure that correlates with subjective test results, which are superior over the final decision as to the quality of speech.

One way to create an objective measure is by separately checking the altered acoustic properties of a signal. An example of using acoustic descriptors is presented in work by Valentini-Botinhao *et al.* [4], where they determined which modifications have a significant impact on the intelligibility of synthetic speech in noise. In this research, we use a feature vector containing acoustic parameters to evaluate the Lombard effect suggested in our previous studies [5]. We compared the performance of this measure with a standardized objective indicator as well as subjective test results.

The Lombard effect is admittedly a long-known phenomenon, discovered in 1911 [6], and from that time on, intensively researched and applied in many areas [7]–[10]. One of the shortest definitions of this effect refers to an

The associate editor coordinating the review of this manuscript and approving it for publication was Baoping Cai.

involuntary increase of vocal response to the presence of background noise. It should, however, be remembered that the Lombard effect modifies not only the volume of the uttered speech, but some other changes also occur such as a fundamental frequency rise, formant frequency rise, spectral tilt, duration of utterances (both elongation and shortening), prosody alteration, etc. [11]–[13]. Moreover, it was also reported that this effect, even though involuntary, may be inhibited and trained in the presence of noise [14]–[16]. The discussion between the spontaneous and inhibited Lombard effect is still ongoing [13], [15]. Since the discovery was related to the audiology domain, it is not surprising that the first applications were related to speech-in-noise audiometry [15]. Interest in employing the Lombard effect in the medical domain also led to improving low voice intensity in Parkinson's disease patients [17], [18], even though applying elevated noise levels in humans for everyday communication seems a challenging concept to be fully approved. Most of both research and application areas are, however, related to human (and human-computer) communication, telecommunications, etc. [9], [19], [20]. Especially important are strategies for improving speech comprehensibility in noisy conditions based on various techniques, including speech modeling. It should be noted that the topic of this article may also contribute to the Quality of Experience, a key factor within telecommunications, Internet speech codec telephony, and speech quality measurement [21], [22].

As mentioned earlier, several features of Lombard speech have been identified in numerous studies, including raising the fundamental frequency or shifting energy from lower frequency bands to medium and higher frequencies. Our work focuses primarily on employing speech models to apply them to Lombard speech without changing the parameters. Our previous investigations have shown that the Lombard speech model, based on dividing the speech signal into harmonics and modeling them as the output of a SISO (Single-Input and Single-Output) system whose transfer function poles are multiple and inputs vary in time, retains Lombard effect characteristics [5]. We explore several speech models, such as sinusoidal, and source-filter applied to modify the speech signal to be compared with our approach. In this study, we also investigate whether the modeled speech is recognizable in unfavorable noise conditions.

Representing a speech signal by a sinusoid with time-varying amplitude and time-varying frequency is a prevalent method in speech modeling. A variety of techniques for synthesis in sinusoidal speech modeling have been proposed by researchers [23], [24]. The broad applicability of the sinusoidal approach is the main reason that this modeling technique is included in our research. In speech recognition studies, the short-phase spectrum is still rather infrequently taken into account. However, some scientists believe that the phase-based representation contributes to speech intelligibility just as much as the corresponding power spectrum [25], [26]. Moreover, Deng *et al.* pointed out that speech emotion recognition and speech enhancement areas may benefit from modifying the short-phase spectrum [26]. That is why, in our work, sinusoidal models without phase preserving, and those with phase preserving are created to compare their efficiency in terms of speech quality measured in noisy conditions. An alternative to the sinusoidal paradigm is the source-filter model. The source-filter model is widely used in synthesizing human speech, as well as musical instrument sounds [27], [28]. We use this model for the research presented here because it is capable of high-quality speech synthesis. Also, the source-filter model is implemented in the most popular vocoders which perform statistical parametric speech synthesis [29]–[33]. However, it should be remembered that the aim of this article is not to compare the implementation effectiveness of vocoders, but to employ speech models for synthesizing Lombard speech in the context of noisy conditions.

The paper is organized as follows; first, the speech modeling techniques are briefly described. Their presentation includes three groups of models, i.e., the harmonic model, source-filter model, and a model based on sinusoids. Then, the overview of the experiments is shown with a block diagram depicting all steps of the analysis performed. Following that, the methods employed for the quality evaluation of the models implemented are shown. They consist of applying objective and subjective measures. To that end, the ITU-T P.563 recommendation and a method proposed by the authors, based on acoustic parameters derived from the speech signal, are applied and compared to modified MUSHRA subjective test results. The ANOVA statistical analysis is then utilized to check the correlation between the objective and subjective test results. In Section 5, the data analyzed are described, and the results of both objective and subjective speech quality evaluations are presented. Then, their statistical consistency is compared. Finally, conclusions are drawn, and the continuation of these studies in the future is outlined.

## II. BACKGROUND ON SPEECH MODELING TECHNIQUES

In this study, the harmonic, source-filter, and sinusoidal models of Lombard speech are investigated. A description of each model is given in this Section.

### A. HARMONIC MODEL

For harmonic modeling, a generator system proposed by Korvel and colleagues [5], [34] is used in this article. The model is based on dividing the speech signal into harmonics and modeling them as the output of a SISO system. The impulse response $h_k\,(n)$ of the system is the $4^{th}$ order quasipolynomial, and is described by the following formula:

$$h_k\,(n) = e^{-\lambda_k n\Delta t} \sum_{m=1}^{4} a_{km}(n\Delta t)^{m-1} sin(2\pi k f_k n\Delta t + \varphi_{km}) \tag{1}$$

where $n$ is the discrete-time, $\Delta t$ is the sampling period, $k = 1, \ldots, K$ ($K$ refers to the number of harmonics), $\lambda_k$ is the

damping factor, $f_k$ is the frequency, and $a_{km}$ and $\varphi_{km}$ are the amplitudes and phases respectively ($m = 1, \ldots, 4$).

The inputs of the $k^{th}$ harmonic system can be described as follows:

$$u_k = [u_{k,1}, u_{k,2}, \ldots, u_{k,L}] \tag{2}$$

where $u_{k,i}$ is the $i^{th}$ input value of the $k^{th}$ harmonic and is calculated as the maximum amplitude of the $i^{th}$ period of the $k^{th}$ harmonic, $L$ is the period number of generated speech signal. The detailed procedure of determining the inputs and the distances between them is presented in a paper by Pyž *et al.* [35].

### B. SOURCE-FILTER MODEL

Based on the source-filter theory, the speech signal is produced by an excitation, which is then filtered by a vocal tract shape. The vocal tract filter can be described as a linear time-invariant system. The mathematical expression of the speech signal model, denoted by $y(t)$, which is an output signal of such a system, is the following:

$$y(t) = h(t) * x(t) \tag{3}$$

where symbol $*$ refers to the convolution operation,

$$h(t) * x(t) = \int_{-\infty}^{\infty} x(\tau)\, h(t-\tau)\, d\tau \tag{4}$$

and $h(t)$ is the impulse response of the system, and $x(t)$ is the input signal.

In this research, to achieve a high-quality speech model, two models based on different architectures are constructed. In both of them, the input signal is a pulse train with a fundamental period.

#### 1) SOURCE-FILTER MODEL WITH APERIODICITY PARAMETER

In the source-filter model, the excitation depends on a fundamental frequency ($f0$); therefore, first, the $f0$ contour is estimated. For this purpose, a method based on both time interval and frequency cues is used [36]. This method provides fundamental frequency and periodicity information within each frequency band. The aperiodicity information is estimated from the residuals between harmonic components and is used to synthesize both the periodic and aperiodic signals. Although in the source-filter theory, the source signal and the vocal-tract filter are separated, under real conditions, there is interaction between them. Therefore, the $f0$ parameter is included in the spectral envelope estimation algorithm. The basic principles of this algorithm can be found in a paper by Kawahara [37]. The algorithm extracts a smoothed time-frequency representation. The reconstructed spectrogram is commonly known as a STRAIGHT spectrogram.

The estimated parameters (STRAIGHT spectrogram, aperiodicity, and $f0$ parameter) are employed for model creation. An overlap-add synthesis using minimum-phase impulse response with group delay manipulations is used for this purpose.

#### 2) SOURCE-FILTER MODEL WITH WAVEFORM-BASED PARAMETER

It is well-known that the human voice is not perfectly periodic. This is why in speech synthesis, a mixed excitation signal containing an aperiodic signal should be applied. According to Morise [38], when a periodic signal is calculated as the minimum-phase response, the model cannot represent the phase of the input voice as the vocal tract response generally includes not only a minimum-phase response but also a maximum-phase response. The author pointed out that to accurately synthesize a voice, it is essential to extract the phase of the input signal.

In this model, an instant of aperiodicity information, waveform-based parameter is used. The model is realized using a high-quality speech analysis, modification, and synthesis system developed by Morise *et al.* [30]. It consists of three analysis algorithms for obtaining speech parameters and one synthesis algorithm that takes these parameters as inputs. In the process of analysis, first of all, the $f0$ parameter and spectral envelope are estimated. As in the case of the source-filter model described above (see Subsection II.B.1), the $f0$ information is also used in the spectral envelope estimation process. The fundamental frequency, spectral envelope information, and the signal waveform, are used for estimation of the excitation signal. During the modeling process, these estimations are incorporated. The details of the algorithm implemented are presented in earlier works by these authors [38]–[40].

### C. SINUSOIDAL MODEL

According to the sinusoidal speech modeling technique, the signal is represented as a sum of sinusoids whose frequencies and amplitudes vary in time. In this research, the parameters of the sinusoids are determined by tracking the spectral peaks, as per the example given in Ellis [41].

#### 1) SINUSOIDAL MODEL WITHOUT PHASE PRESERVING

The construction of the model begins with the construction of the sinusoidal representation of the speech signal. For this purpose, the Short-Time Fourier Transform (STFT) spectrogram, which is a visual representation of the signal spectrum that varies with time, is used.

Let:

$$x = [x(1), x(2), \ldots, x(N)]^T \tag{5}$$

be a sequence of samples of the analyzed speech signal, where $N$ is the number of samples per signal and $[.]^T$ represents the matrix transpose operation.

Signal $x$ (see Eq. (5)) is divided into short-time segments with overlaps between adjacent segments equal to 50%, and each segment is windowed with a Hamming window. The length of a segment is equal to 512 points. The magnitude spectrum of the $l$-th short-time segment (denoted by $x_l$) is obtained by the following formula:

$$|X_l(k)| = \frac{1}{M_{FT}} \sqrt{(X_l(k))_{re}^2 + (X_l(k))_{im}^2} \tag{6}$$

where $X_l(k)$ is the Fourier transform of the short-time segment $x_l, k = 1, \ldots, M_{FT}$ ($M_{FT}$ refers to the number of Fourier transform coefficients), $l = 1, \ldots, L$ ($L$ refers to the number of short-time segments).

Based on the spectrogram, a speech analysis is performed, which determines the stationary and deterministic parts of the speech signal. For this purpose, frequencies and amplitudes corresponding to local peaks in the spectrum are detected. The other task is to determine which peaks belong to the spoken signal. To achieve this, the list of detected peaks is fed into a tracking procedure. According to this procedure, for each frequency $\omega_i^k$ in frame $k$ we are looking for the frequency $\omega_j^{k+1}$ in frame $k+1$ is sought, which is closest to such a frequency and whose absolute distance is less than the threshold $\Delta$, i.e.:

$$|\omega_i^k - \omega_j^{k+1}| < |\omega_i^k - \omega_p^{k+1}| < \Delta \tag{7}$$

where the $|.|$ symbol refers to the absolute value or magnitude, $i = 1, \ldots, L_k$, ($L_k$ – the total number of peaks in frame $k$), $j = 1, \ldots, L_{k+1}$, ($L_{k+1}$ – the total number of peaks in frame $k+1$, and $(p = 1, \ldots, L_{k+1}) \cap (p \neq j)$.

If no match between frequencies is found, they are matched to themselves, and their magnitudes are set to zero. As a result, we obtain an interpolated peak magnitude for each track point.

In the last step, speech signal resynthesis is performed. For reconstruction, a sine wave oscillator bank developed by Ellis is used [41].

#### 2) SINUSOIDAL MODEL WITH PHASE PRESERVING

Most speech processing applications are based on the short-time spectrum, while relatively little attention is paid to the short-range phase spectrum. According to Abe and colleagues, the Instantaneous Frequency (IF) spectrogram more clearly shows the harmonic structure of quasi-periodic signals such as speech than STFT spectrograms [42]. The advantages of including phase-related information in the speech vocoder are listed in these works [43], [44].

In this model, instead of the STFT spectrogram, which discards phase information, the IF spectrogram is used. The harmonic frequencies based on the IF of a speech signal are obtained by the technique proposed in work by Abe and colleagues [45].

Resynthesizing consists of reading the series of frequency, magnitude, and phase samples for a particular track. For this purpose, the Matlab code developed by Ellis [41] is utilized.

## III. EXPERIMENTAL SETUP

The main goal of the experiment was two-fold: first, to check the level of recognizability of the speech models with the applied Lombard effect, and to determine at what noise threshold a particular model stops working. To that end, a quality measure was introduced, based on a feature vector derived from the signal analyzed, and then compared with the standardized metric (as described in Section IV) as well as with the MUSHRA test results. For this purpose, the models given in Section II were created utilizing all recorded speech utterances with the Lombard effect. The block diagram of the experimental setup is presented in Fig. 1.

We use the following denotations of the speech models:

M1 – harmonic model,

M2 – source-filter model with an aperiodicity parameter,

M3 – source-filter model with a waveform-based parameter,

M4 – sinusoidal model without phase preserving,

M5 – sinusoidal model with phase preserving.

Respectively, the denotations for the real speech signals are the follows:

LS – utterance with the Lombard effect,

NS – original, natural speech utterance (non-Lombard).

The experiment consisted of two parts. In the first part, an objective evaluation of the models was performed. The models, as well as the real speech signals, were mixed with babble speech and street noise recordings. Samples of noise were taken from the YouTube platform. The following signal-to-noise ratios (SNRs) were tested: −20 dB, −15 dB, −10 dB, −5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB.

## IV. QUALITY EVALUATION TECHNIQUES APPLIED

The quality of the models created was measured, employing both objective and subjective measures. In this research, two objective indicators, being P.563, defined by ITU-T recommendation [1], and a method based on acoustic parameters, proposed by the authors, are employed. The subjective quality evaluation is obtained by the method described in the ITU-R BS. 1534-1 standard [2], known as MUSHRA (MUltiple Stimulus with Hidden Reference and Anchors).

We decided to use P.563 metrics to calculate the speech quality. It should, however, be remembered that P.563 is a single-ended measure that does not require the source (original) signal to compare. In contrast, double-ended measures are based on the comparison of the original and the degraded signal. In the case of our study, we have the original signal, so double-ended measures – such as PESQ (Perceptual Evaluation of Speech Quality) [46]–[48] – could have been be used. However, the applicability of double-ended measures is limited in the context of our investigations as they will not return accurate value metrics. Assuming that for one type of noise and one type of speech modification, there are four recordings to be evaluated (i.e., without noise and modification, without noise and with modification, with noise and without modification, with noise and with modification), then there are two possible ways of performing PESQ comparisons:

1) The first case refers to the situation in which there is the original signal, without noise or modification, and the degraded signal, which contains noise, and the modification is applied – the PESQ algorithm will treat the modification of the speech signal as degradation, which does not suit the aim of our work.
2) The second case includes the original signal with or without modification, and the degraded signal is with or

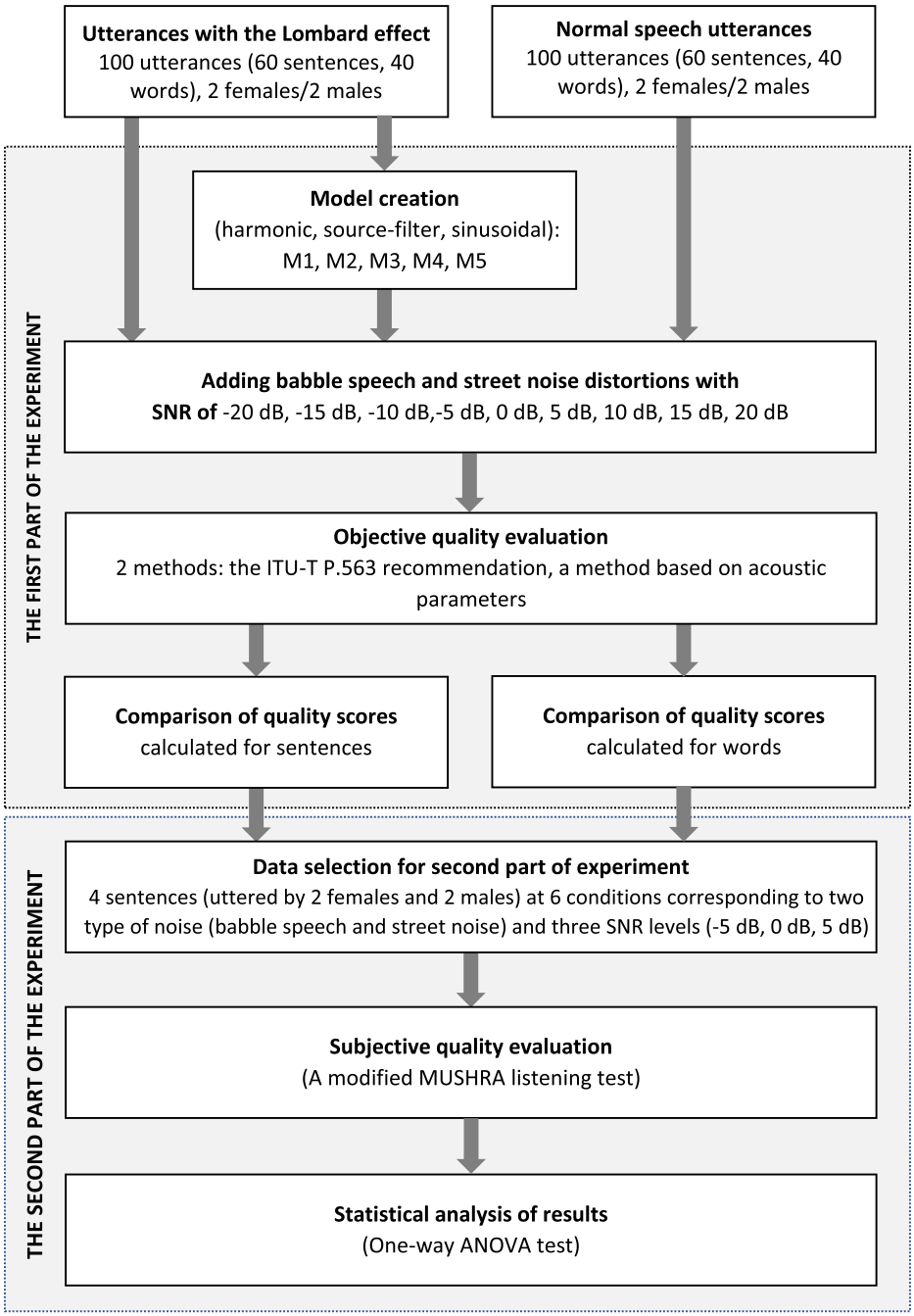


**FIGURE 1.** Block diagram of the experimental setup.

without modification, respectively, still, also with noise applied. Under such conditions, speech modification is not treated as degradation. The metric will show only what the impact of the noise on the speech quality is. But we will not get information on how the modification potentially impacts the speech quality – which is the factor that we would like to measure.

The above consideration shows why double-ended measures cannot be applied to calculate the impact of modifications on the speech quality measured in noisy conditions.

### A. OBJECTIVE QUALITY LEVEL INDICATORS

#### 1) ITU-T RECOMMENDATION P.563

One of the frequently used non-intrusive speech quality measures is defined by the ITU-T P.563 standard [1]. This measure is most often applied in telecommunications because it performs a single-ended verification of the channel quality. This often enables quick and reliable system adaptation, taking the channel quality into consideration. As an output, P.563 measurements return a Mean Opinion Score - Listening Quality Objective (MOS-LQO), which shows quite a high

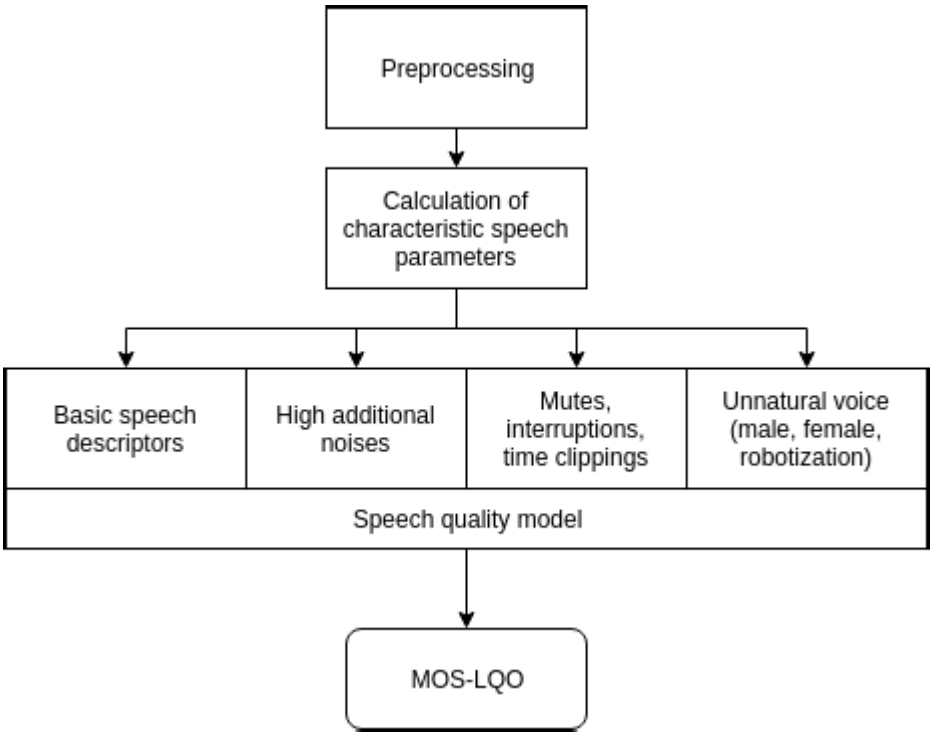


**FIGURE 2.** The basic block scheme of the P.563 algorithm [1].

correlation with the Mean Opinion Score- Listening Quality Subjective (MOS-LQS) values returned from the subjective tests.

Each signal subjected to MOS measurement using P.563 must be pre-processed by using the model of the listening device. In the next stage, a speech detector (VAD – Voice Activity Detector) is used to mark the speech-related signal fragments. Further on, the speech signal is subjected to a series of analyses and assigned to a given class of disturbances. Parameterization of the signal in P.563 can be divided into three basic functional blocks (ITU-T Recommendation P.563 [1]) that correspond to the main classes of distortion:

- analysis of the vocal tract and speech unnaturalness; in this case, it is possible to discern speech defects separately for female and male voices and also identify the so-called ''robot'' effect,
- analysis of strong additional noise; in this case, it is vital to detect the static background noise floor, and noise associated with the signal envelope,
- analysis of interruptions, mutes, time clipping, and cuts.

The basic block scheme of the P.563 algorithm [1] is shown in Fig. 2.

In single-sided measurements, the MOS value is estimated exclusively on the basis of the interference signal. In the case of the P.563 standard, the use of a real expert listening to the conversation on a test device should be simulated. This device can be any receiver, e.g., a mobile phone. Since, in this case, the degraded signal is not compared to the original signal, the speech quality indicator depends on the listening device. It is, therefore, an important element of the P.563 standard (ITU-T Recommendation P.563 [1]).

The test signal must also meet the requirements specified in the standard, so that it is possible to detect the speech quality using the P.563 algorithm, including:

- the sampling frequency must be greater than or equal to 8 kHz,
- the digital signal resolution must be 16-bit,
- the signal cannot be longer than 20 seconds, and the speech content in the signal cannot be shorter than 3 seconds.

Some research studies show a high correlation between the MOS-LQS and MOS-LQO obtained with the P.563 algorithm [49], [50], but different speech characteristics can be evaluated here, for instance, speech naturalness or intelligibility. P.563 measurement correlations with subjective tests also depend on the speech sampling frequency – comparisons show that a 16 kHz sampling frequency provides better correlation in terms of naturalness and intelligibility [50]. This is why the authors of this article used speech recordings resampled to 16 kHz as an input of the objective measurements.

### 2) SPEECH QUALITY INDICATOR BASED ON ACOUSTIC PARAMETERS

In our previous research related to Lombard speech models, we suggested a set of parameters that let us evaluate the Lombard effect in speech [5]. In this research, we propose a measure of model quality, depending on these parameters. The list of the signal descriptors is given in Table 1.

**TABLE 1. Acoustic parameters for evaluation of the Lombard effect in models [5].**

| | |
|---|---|
| The time-domain parameters | |
| 1 | Temporal Centroid |
| 2 | Zero Crossing Rate |
| 3 | Root Mean Square (RMS) energy |
| 4-6 | The number of samples exceeding levels RMS, 2×RMS, 3×RMS |
| 7-12 | The mean and variance of samples exceeding levels RMS, 2×RMS, 3×RMS averaged for 10 sub-segments |
| 13 | Peak to RMS |
| 14-17 | The number of the signal crossings in relation to zero, RMS, 2×RMS, 3×RMS |
| 18-25 | The mean and variance of signal crossings in relation to zero, RMS, 2×RMS, 3×RMS averaged for 10 sub-segments |
| The frequency-domain parameters | |
| 26-30 | The first five formants |
| 31 | Audio Spectral Centroid |
| 32 | Audio Spectral Spread |
| 33 | Audio Spectral Skewness |
| 34 | Audio Spectral Kurtosis |
| 35 | Spectral Entropy |
| 36 | Spectral Roll-Off |
| 37 | Spectral Brightness |
| 38-66 | Audio Spectrum Envelope calculated on 29 sub-bands |
| 67 | Mean Audio Spectrum Envelope |
| 68-85 | Spectral Flatness Measure calculated on 18 sub-bands |
| 86 | Mean Spectral Flatness Measure |
| 87-106 | Mel-Frequency Cepstral Coefficients |

The parameters (see Table 1) include time and frequency domain features. The frequency-domain parameters are calculated from the Fourier spectrum. The speech signal is divided into short-time segments with a 50% overlap, and each segment is windowed with the Hamming window before the parameters are calculated. A more precise description of the parameters listed in Table 1 is given in the authors' publications [51]–[53].

The measure of model quality is based on the normalized distances between the parameters. The distances are described by the following formula:

$$Dist = \sum_{i=1}^{N} \frac{\mid SD_i\,(Lomb) - SD_i(model) \mid}{Max\_par_i} \tag{8}$$

where $N$ is the number (i.e., $N = 106$) of parameters, $SD_i\,(lomb)$ is the standard deviation of the $i^{th}$ Lombard speech feature vector calculated on short-time segments, and $SD_i(model)$, the standard deviation of the $i^{th}$ feature vector of the model derived from short-time segments, is calculated as follows:

$$SD_i = \sqrt{\frac{\sum_{j=1}^{M} (r_{ij} - \overline{r_i})^2}{M - 1}} \tag{9}$$

where $M$ is the number of short-time segments, $r_{ij}$ – the $j^{th}$ value of the $i^{th}$ feature vector, $\overline{r_i}$ – the mean value of the $i^{th}$ feature vector.

The $Max\_par_i$ parameter is calculated as the maximum value of the $i^{th}$ feature vector of the natural speech, i.e.:

$$Max\_par_i = max\{r_{ij}\,(Lomb)\} \tag{10}$$

In the last step of measure construction, the distances (see Eq. (8)) are normalized to the interval [1, 5], which corresponds to the MOS-LQS scale.

### B. SUBJECTIVE QUALITY EVALUATION

When it comes to speech models, a subjective test is an essential element of the evaluation process, which allows the quality of the obtained sounds to be assessed. Therefore, a subjective evaluation of the speech models is also included in this article. This evaluation is based on a modified MUSHRA listening test. The modification applied will be explained later on.

MUSHRA stands for MUlti Stimulus test with Hidden Reference and Anchor. It is a test that performs a subjective comparison of multiple audio signals, and it is suitable for intermediate audio quality [2]. MUSHRA is described in ITU recommendation BS.1534-1 [2] and updated in BS.1534-2 [54].

There are some requirements that describe the MUSHRA test, for instance:

1) The sequence should not exceed 20 seconds to prevent the listeners' fatigue and to reduce the total duration of the listening test;
2) In total, a session should not last for more than 20 minutes to avoid fatigue in judgments;
3) The set of signals should contain one reference signal (full quality) and one low-pass filtered signal version (the so-called anchor, typically with 3.5 kHz bandwidth); additional anchors might be used optionally.

Despite the usability of MUSHRA, one should be aware of potential biases that may occur when preparing test signals and constructing the whole set to be evaluated [52]. In the designed experiments, Zieliński showed systematic discrepancies in the results in the MUSHRA test [55]. They refer to stimulus spacing bias, centering bias, range equalizing bias, contraction bias, and bias due to nonlinear properties of an assessment scale. The possible biases that may occur in the tests performed will be discussed in the Conclusion Section.

The authors of this work created the MUSHRA test using the web interface and the Audiolabs' MUSHRA application [56]. It was installed on a web server and configured using the following assumption: every page in the MUSHRA test contains a single sentence of a single person with different types of modifications with the same level and type of noise.

The test is available online. It was, however, modified to it adapt to the quality of the presented signals. Test users in the pre-test session reported that the clean (reference) signal and the low-pass filtered anchors disturbed the overall listening experience during the test, thus not allowing for the correct quality assignment. The authors, therefore, removed the reference signal and the anchor. That is why, in this work, the test is referred to as a ''modified MUSHRA test''.

For the statistical analysis of the data obtained through the MUSHRA method, the ANOVA (Analysis of Variance) test, which is supported by the recommendation, was used [2].

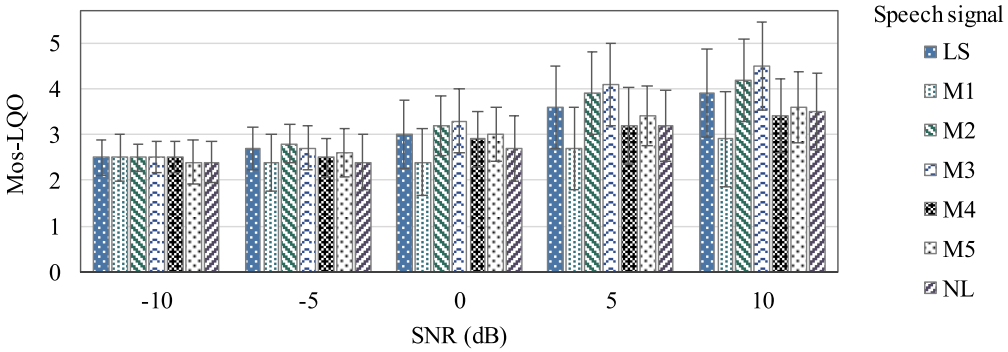


**FIGURE 3.** Estimated averaged MOS-LQO values for babble speech distortions (calculated for recordings containing sentences). Denotations are as follows: speech models: M1 – harmonic model, M2 – source-filter model with aperiodicity parameter, M3 – source-filter model with a waveform-based parameter, M4 – sinusoidal model without phase preserving, M5 – sinusoidal model with phase preserving; real speech signals: LS – utterance with the Lombard effect, NS – original, natural speech utterance.

## V. EXPERIMENT RESULTS

### A. DATA ANALYZED

The experiments are performed on recordings of four speakers (two males and two females). The speakers were asked to read 25 statements, which included 15 sentences in Polish with different prosody (indicative, imperative, and questioning utterances) and 10 separate words. The sentences and words used in the experiment are listed in the Appendix. These statements were recorded in.wav audio files with the following parameters: 48 kHz; 16 bit; mono. The recording of utterances was carried out in a room with an acoustically treated interior which suppresses reverberation. The recording procedure was repeated twice: without additional noise as well as with noise interference. To simulate noise conditions, closed headphones were used. As a result, two types of recordings: 100 statements of natural, normal speech, i.e., non-Lombard speech and 100 with the Lombard effect, were obtained.

### B. RESULT ANALYSIS

The objective evaluation of the recordings was performed separately for words and sentences. The obtained results for the sentences are given in Tables 2-3. Scores rated the same or higher in comparison with Lombard speech (LS) are highlighted in bold font.

A graphical representation of the results given in Tables 2-3 for babble speech noise is presented in Figs. 3-4. A graphical representation of the results given in Tables 2-3 for street noise is presented in Figs. 5-6.

When referring to the speech-in-noise conditions, typically, speech utterances are analyzed in the context of the occurrence of the Lombard effect. However, in this work, separated words were also tested to see if the Lombard effect could be applied to a single word, and if it could have an impact on speech quality. The obtained results for recordings containing words are given in Tables 4-5, where the scores rated the same or higher in comparison with Lombard speech (LS) are highlighted in bold font.

**TABLE 2.** Estimated averaged MOS-LQO values for babble speech and street noise distortions (recordings containing only sentences, not words were used in the evaluation process).

| Babble speech noise | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SNR | –20 dB | –15 dB | –10 dB | –5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| LS | 2.43 | 2.47 | 2.52 | 2.72 | 3.01 | 3.61 | 3.91 | 3.97 | 4.17 |
| M1 | **2.45** | 2.43 | 2.46 | 2.39 | 2.43 | 2.71 | 2.85 | 2.90 | 3.55 |
| M2 | **2.43** | 2.46 | 2.48 | **2.75** | **3.21** | **3.87** | **4.23** | **4.32** | **4.44** |
| M3 | 2.38 | 2.39 | 2.46 | **2.72** | **3.28** | **4.06** | **4.45** | **4.64** | **4.69** |
| M4 | 2.41 | 2.44 | 2.46 | 2.54 | 2.91 | 3.20 | 3.36 | 3.60 | **4.21** |
| M5 | 2.37 | 2.41 | 2.41 | 2.56 | 2.99 | 3.42 | 3.59 | 3.66 | 4.16 |
| NL | 2.41 | 2.46 | 2.42 | 2.44 | 2.67 | 3.21 | 3.49 | 3.65 | 3.70 |
| Street noise | | | | | | | | | |
| SNR | –20 dB | –15 dB | –10 dB | –5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| LS | 1 | 1 | 1 | 1 | 1.28 | 2.70 | 3.70 | 3.49 | 3.93 |
| M1 | 1 | 1 | 1 | 1 | 1.16 | 2.38 | 3.16 | 3.09 | 3.48 |
| M2 | 1 | 1 | 1 | 1 | 1.23 | **2.78** | **3.76** | **3.89** | **4.00** |
| M3 | 1 | 1 | 1 | 1 | **1.32** | **2.78** | 3.68 | **3.76** | **4.04** |
| M4 | 1 | 1 | 1 | 1 | **1.30** | 2.66 | 3.63 | 3.31 | 3.84 |
| M5 | 1 | 1 | 1 | 1 | **1.38** | **2.88** | **3.86** | **3.76** | **3.93** |
| NL | 1 | 1 | 1 | 1 | 1.23 | 2.32 | 3.10 | 3.23 | 3.66 |

**TABLE 3.** Estimated averaged quality scores for babble speech and street noise distortions obtained by the method based on acoustic parameters derived from speech (recordings containing only sentences, not words were used in the evaluation process).

| Babble speech noise | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SNR | –20 dB | –15 dB | –10 dB | –5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| LS | 1.13 | 1.18 | 1.52 | 2.62 | 3.58 | 4.59 | 4.92 | 4.99 | 4.98 |
| M1 | **1.29** | **1.33** | **1.54** | 2.56 | 3.50 | 4.37 | 4.62 | 4.65 | 4.62 |
| M2 | 1.12 | **1.19** | 1.52 | **2.87** | **3.82** | **4.69** | **4.95** | **5.00** | **4.99** |
| M3 | **1.15** | **1.18** | 1.46 | **2.75** | **3.72** | **4.64** | **4.94** | **4.99** | **4.98** |
| M4 | 1.12 | **1.18** | 1.33 | 2.21 | 3.02 | 4.22 | 4.59 | 4.63 | 4.59 |
| M5 | **1.16** | **1.21** | 1.49 | **2.65** | **3.61** | **4.61** | **4.92** | 4.98 | 4.96 |
| NL | 1 | 1.04 | 1.25 | 2.38 | 3.37 | 4.44 | 4.78 | 4.86 | 4.85 |
| Street noise | | | | | | | | | |
| SNR | –20 dB | –15 dB | –10 dB | –5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| LS | 1 | 1.06 | 1.22 | 2.18 | 3.91 | 4.74 | 4.96 | 4.98 | 4.96 |
| M1 | 1 | 1.05 | 1.17 | 2.12 | 3.64 | 4.40 | 4.58 | 4.58 | 4.56 |
| M2 | 1 | 1 | 1.19 | **2.49** | **4.13** | **4.83** | **5.00** | **5.00** | **4.98** |
| M3 | 1 | 1 | 1.18 | **2.39** | **4.03** | **4.79** | **4.98** | **4.99** | **4.97** |
| M4 | 1 | 1.05 | 1.16 | 1.88 | 3.48 | 4.38 | 4.62 | 4.60 | 4.53 |
| M5 | 1 | 1 | **1.21** | **2.20** | **3.95** | **4.77** | **4.97** | **4.98** | **4.94** |
| NL | 1 | 1 | 1.18 | **2.26** | 3.89 | 4.66 | 4.84 | 4.84 | 4.82 |

Since the results obtained for words are partly similar to the results obtained for sentences, we do not show their graphical representations.

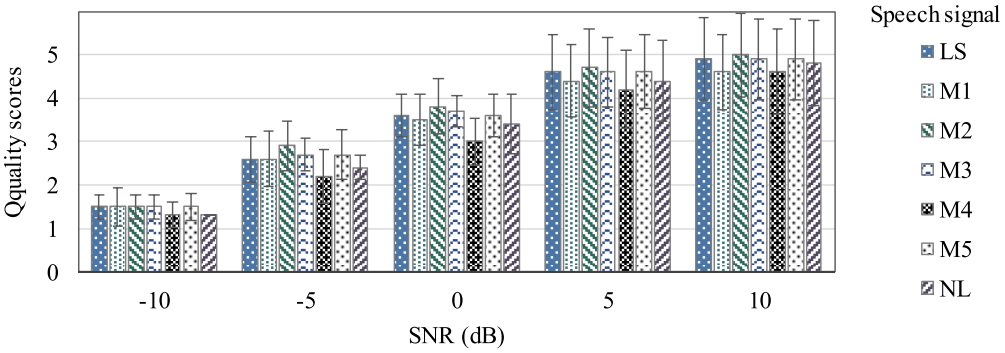


**FIGURE 4.** Estimated averaged quality scores for babble speech distortions obtained by the method based on acoustic parameters (calculated for recordings containing sentences); denotations as shown in Fig. 3.

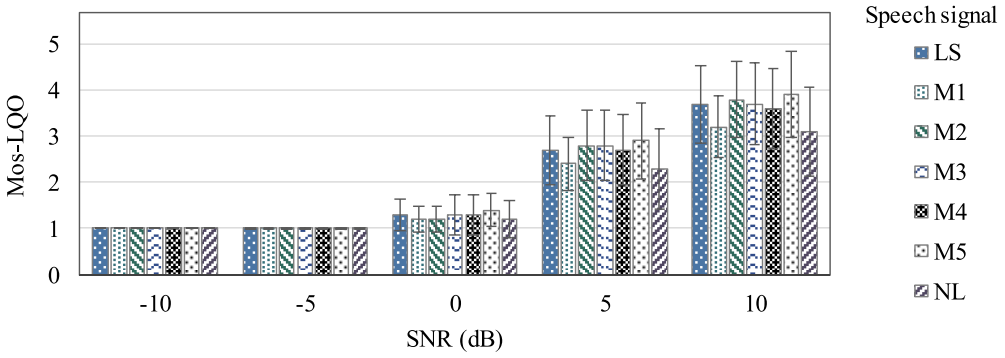


**FIGURE 5.** Estimated averaged MOS-LQO values for street noise distortions (calculated for recordings containing sentences); denotations as shown in Fig. 3.

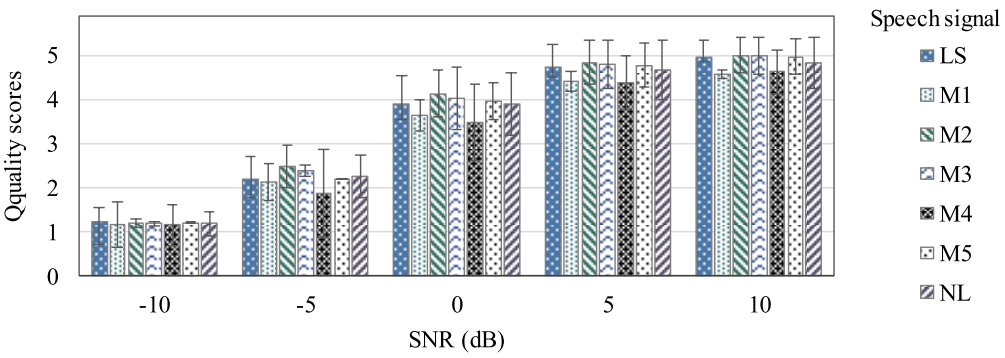


**FIGURE 6.** Estimated averaged quality scores for street noise distortions obtained by the method based on acoustic parameters (calculated for recordings containing sentences); denotations as shown in Fig. 3.

According to the grading scale, the quality of sounds which achieved an approximated score equal to 3, which refers to ''fair'' quality, may be considered as slightly annoying. Based on this result, the answer may be given as to at what SNR level threshold a particular model stops working. The results indicate that this threshold is −5 dB in the case of babble speech noise for both objective quality evaluation techniques. For street noise distortions, thresholds are 5 dB and 0 dB with respect to the P.563 indicator and the method based on acoustic parameters, respectively.

Based on the objective results of the word model evaluation (see Tables 4-5), the same SNR level thresholds were established as in the case of the sentence model assessment (i.e., −5 dB in the case of babble speech noise, in addition to 5 dB and 0 dB for street noise distortions).

It is worth noting that with the addition of babble noise of a very high volume, the MOS values at SNRs at −20 dB, −15 dB, and −10 dB indicate that the sound quality is good enough (see Tables 2, 4). However, these results are not reliable because the estimated LS values are lower than the NS values. In fact, the opposite is true, i.e., the LS values at high noise levels should be higher than those of the NS. This may be caused by the fact that the added babble noise contains speech, and the quality ratings obtained refer to the noise rather than to the signal.

The objective measures show that in most cases, the best scores are achieved with both source-filter models and the

**TABLE 4.** Estimated averaged MOS-LQO values for babble speech and street noise distortions (recordings containing only words were used in this part of the evaluation process).

| Babble speech noise | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SNR | –20 dB | –15 dB | –10 dB | –5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| LS | 2.27 | 2.46 | 2.65 | 3.01 | 3.63 | 4.16 | 4.14 | 4.26 | 4.28 |
| M1 | **2.29** | **2.47** | 2.58 | 2.66 | 3.15 | 3.30 | 3.35 | 3.63 | 3.71 |
| M2 | **2.30** | 2.44 | 2.64 | 2.97 | **3.77** | **4.35** | **4.46** | **4.40** | **4.44** |
| M3 | **2.41** | 2.44 | 2.58 | **3.07** | **3.83** | **4.39** | **4.54** | **4.55** | **4.54** |
| M4 | **2.32** | **2.52** | 2.54 | 2.91 | 3.55 | 3.74 | 3.82 | 3.92 | 3.92 |
| M5 | **2.29** | **2.48** | **2.65** | 2.85 | 3.51 | 3.87 | 3.70 | 3.83 | 3.73 |
| NL | **2.47** | **2.53** | **2.68** | 2.72 | 3.12 | 3.55 | 3.81 | 3.88 | 4.10 |
| Street noise | | | | | | | | | |
| SNR | –20 dB | –15 dB | –10 dB | –5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| LS | 1.35 | 1.36 | 1.38 | 1.24 | 1.90 | 3.40 | 4.06 | 4.02 | 4.15 |
| M1 | **1.40** | **1.41** | **1.43** | **1.33** | 1.78 | 3.08 | 3.71 | 3.75 | 3.87 |
| M2 | 1.34 | 1.35 | 1.37 | 1.23 | **1.97** | **3.67** | **4.18** | **4.05** | **4.21** |
| M3 | 1.33 | **1.36** | 1.36 | 1.23 | **1.99** | **3.71** | **4.18** | **4.16** | **4.39** |
| M4 | 1.34 | **1.36** | **1.40** | 1.21 | **2.05** | 3.34 | **4.15** | **4.11** | **4.29** |
| M5 | **1.35** | 1.35 | 1.37 | 1.24 | **2.00** | **3.43** | 3.82 | 3.70 | 3.80 |
| NL | **1.53** | **1.53** | **1.54** | **1.48** | 1.81 | 3.00 | 3.52 | 3.54 | 3.92 |

**TABLE 5.** Estimated averaged quality scores for babble speech and street noise distortions obtained by the method based on acoustic parameters (recordings containing only words were used in this part of the evaluation process).

| Babble speech noise | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| SNR | –20 dB | –15 dB | –10 dB | –5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| LS | 1.29 | 1.32 | 1.60 | 2.44 | 3.12 | 4.34 | 4.82 | 4.96 | 4.99 |
| M1 | 1.23 | 1.24 | 1.47 | 2.18 | 2.70 | 3.82 | 4.14 | 4.21 | 4.20 |
| M2 | **1.32** | **1.37** | **1.70** | **2.49** | **3.49** | **4.45** | **4.86** | **4.97** | **5.00** |
| M3 | 1.28 | 1.30 | **1.68** | **2.52** | **3.33** | **4.41** | **4.83** | 4.92 | 4.93 |
| M4 | 1.28 | 1.28 | 1.41 | 2.12 | 2.57 | 3.77 | 4.19 | 4.23 | 4.19 |
| M5 | 1.26 | 1.23 | 1.40 | 1.95 | 2.12 | 3.30 | 3.73 | 3.80 | 3.78 |
| NL | 1.00 | 1.07 | 1.35 | 2.37 | **3.17** | 4.12 | 4.43 | 4.50 | 4.50 |
| Street noise | | | | | | | | | |
| SNR | –20 dB | –15 dB | –10 dB | –5 dB | 0 dB | 5 dB | 10 dB | 15 dB | 20 dB |
| LS | 1 | 1 | 1 | 2 | 3.62 | 4.55 | 4.89 | 4.97 | 4.99 |
| M1 | 1 | 1 | 1 | 2 | 3.40 | 4.24 | 4.53 | 4.57 | 4.57 |
| M2 | 1 | 1 | 1 | 2 | **3.76** | **4.62** | **4.93** | **5.00** | **5.00** |
| M3 | 1 | 1 | 1 | 2 | **3.74** | **4.61** | **4.90** | 4.96 | 4.96 |
| M4 | 1 | 1 | 1 | 2 | 3.21 | 4.18 | 4.52 | 4.57 | 4.55 |
| M5 | 1 | 1 | 1 | 2 | 2.98 | 3.91 | 4.30 | 4.37 | 4.36 |
| NL | 1 | 1 | 2 | 2 | **3.77** | 4.49 | 4.71 | 4.74 | 4.73 |

model based on the sinusoids with phase preserving. In contrast, the measure based on parameterization shows a smaller difference between the models than in the case of the P.563 indicator values. A listening test should be performed to check whether the proposed measure overestimates the quality of the models, or the MOS-LQO underestimates it.

When comparing the obtained results to the state-of-the-art, one can see that such a comparison in practice is not straightforward. For example, Michelsanti *et al.* [57] reported averaged scores of PESQ and ESTOI (Extended Short-Time Objective Intelligibility) [58] measures for a deep-learning-based system of audio-visual speech enhancement with the Lombard effect applied. To elicit the Lombard effect, Speech Shaped Noise (SSN) at 80 dB Sound Pressure Level (SPL) was presented to the speakers while they were reading the sentences [57]. It is worth noting that ESTOI scores, which estimate speech intelligibility, range from 0 to 1, where high values correspond to high speech intelligibility. When trained on a narrow SNR range, for the audio-only case with the Lombard effect (AO-L), the PESQ measurement returned a value of 1.283, and the ESTOI was equal to 0.448. Contrarily,

**TABLE 6.** Subjective quality scores for babble speech and street noise distortions; denotations are as follows: real speech signals: LS – utterance with the Lombard effect, speech models: M0 s – source-filter model with aperiodicity parameter, M3 – source-filter model with a waveform-based parameter, M5 – sinusoidal model with phase preserving.

| | Babble speech noise | | | Street noise | | |
|---|---|---|---|---|---|---|
| SNR | –5 dB | 0 dB | 5 dB | –5 dB | 0 dB | 5 dB |
| LS | 64.11 | 66.91 | 76.34 | 28.39 | 48.55 | 64.89 |
| M2 | 64.05 | 65.27 | 70.39 | 31.43 | 45.07 | 61.34 |
| M3 | 62.75 | 63.39 | 66.32 | **28.41** | 43.73 | 60.68 |
| M5 | 52.61 | 50.77 | 53.77 | 31.59 | 46.48 | 60.20 |

when the system was trained on a wide SNR range, the averaged values were ranged between 1.346 (for −20 to 5 dB) to 3.127 (for 10 to −30 dB) for the AO-L case. The ESTOI values changed dramatically from 0.442 for −20 dB to −5 dB SNR, up to 0.927 for a SNR range between 10 and 30 dB. So, the relative performance of the systems at SNR $\leq$ 5 dB is similar to that observed for the systems trained on a narrow SNR range [57].

As seen from this discussion, not only the experimental setup was different compared to our approach, but the analysis also differs from that performed by us; thus a direct comparison is not possible. Even the observation of what SNR value the model does not work at seems to be uncommon; in the case of our research, the models stop working at a threshold of −5 dB, in the work of Michelsanti *et al.* [57], it refers to 5 dB.

### C. SUBJECTIVE TEST RESULTS

An informal listening test showed that the quality of Lombard speech models can be directly compared to the original sound. Therefore, in the second part of the experiment, the subjective quality evaluation was carried out. In the listening experiment, the participants compared the performance of different models to the natural utterances of Lombard speech in noise. To ensure that the test session did not take more than 20 minutes, four Lombard speech utterances consisting of sentences uttered by four speakers were used. Also, only the three models, which showed the best results in the first part of the experiment (M2, M3, and M5) and Lombard speech utterances, were evaluated. Because of the time constraint, only two types of noise recordings (babble speech and street noise) were mixed with the speech models. In addition to the above, the following SNRs were considered: −5 dB, 0 dB, and 5 dB. As a result, six test conditions corresponding to combinations of noise types and SNRs were used in the listening test. The average duration of the MUSHRA test session was approx. 20 minutes. Twelve speech processing experts from the Vilnius University Institute of Data Science and Digital Technologies, and the Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics took part in this test. The obtained results are given in Table 6.

The subjective test results show (see Table 6) that original recordings of Lombard speech are more intelligible in under noise conditions than their models (except for one example,

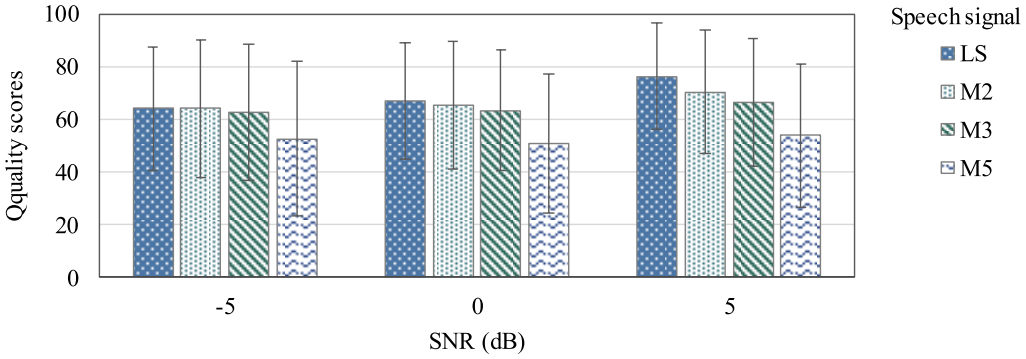


**FIGURE 7.** Subjective quality scores for babble speech noise; denotations as shown in Table 6.

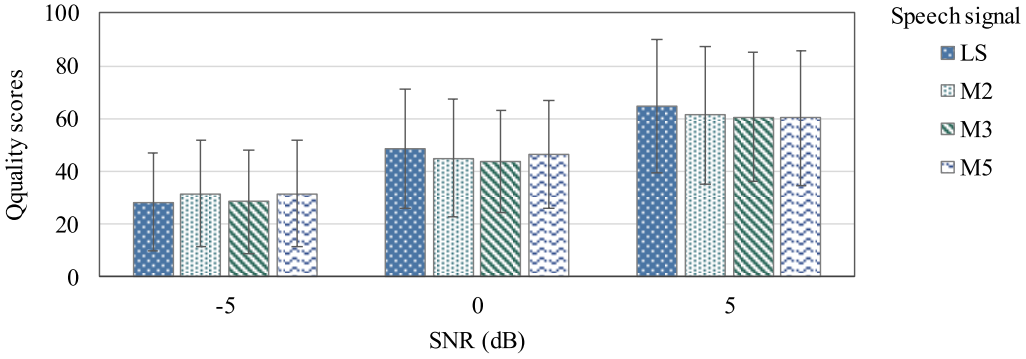


**FIGURE 8.** Subjective quality scores for street noise; denotations as shown in Table 6.

which is highlighted in bold font). A visualization of the subjective test results is given in Figs. 7-8.

In line with the results shown earlier (see Figs. 7–8), one can observe that the SNR level threshold at which a particular model stops working is 0 dB in the case of street noise. In contrast, for babble speech noise, it is not possible to determine such a threshold based on the experiment carried out.

Results obtained by Michelsanti *et al.* [57] refer to two types of subjective tests, namely the MUSHRA and speech intelligibility tests. For an AO-L at −5 dB SNR, the result was approx. 25 points, whereas for 5 dB SNR, results returned approx. 50 points. Obviously, the intelligibility test also depends on the SNR. Moreover, it was tested for several types of words, i.e., color, letter, and digit. The mean intelligibility scores are within the range of approx. 35% for a −20 dB SNR to approx. 85% for a 5 dB SNR. Again, all analysis conditions differ from those used by us, thus a straightforward comparison is not possible. However, the MUSHRA scores for street noise are low for a −5 dB SNR, and they are at the same level as in work by Michelsanti *et al.* [57]. In contrast, they are higher for a SNR equal to 5 dB for both street noise and babble speech conditions.

Seshadri *et al.* [10] reported MUSHRA-based scores when applying the Lombard effect to several vocoders. In order to induce Lombard speech, background noise in the form of nonstationary pub noise, with an A-weighted SPL of approximately 80 dB, was presented to the speakers' ears with headphones while they were being recorded [10]. Scores were shown for parametric vocoders (VOCs) for feature extraction and Machine Learning Models (MLMs) for speech modifications. The MUSHRA test aimed at evaluating the Lombardness of the utterances from different VOC and MLM combinations of a single sentence (same speaker and linguistic content). All results were conditioned by the various vocoders employed. The scores ranged between approx. 40 to 60 points of the mean Lombardness. Moreover, the CMOS (Comparison Mean Opinion Score) quality test was applied as well as the so-called instrumental intelligibility test, given in bits/s. Also, in this case, it is not possible to directly compare the results reported by Seshadri *et al.* and the results obtained in our study.

Lopez *et al.* [59] conducted two subjective tests on a glottal vocoder and the STRAIGHT vocoder compared to natural Lombard speech. The first listening session consisted in evaluating to what extent the vocoder speech samples resembled natural Lombard speech on a continuous scale from 1 (none) to 5 (very much). A pairwise comparison test, aimed at evaluating the naturalness of the converted vocoder Lombard speech samples, was used in the second subjective session. The listeners were asked to indicate which of the vocoder Lombard speech samples sounded more natural [59]. In the results, the authors reported that glottal speech samples were evaluated as better ones, however, with a larger median rate in the case of the male speakers. The median was at the level of 2 and 3 scores, translating into 'little' and 'moderately'

**TABLE 7.** Result of the ANOVA test (*F*-values) for MOS-LQO quality scores.

| | Babble speech noise | | | Street noise | | |
|---|---|---|---|---|---|---|
| SNR | –5 dB | 0 dB | 5 dB | –5 dB | 0 dB | 5 dB |
| M2 | 0.1306 | 2.2696 | 2.8214 | – | 0.6220 | 0.3131 |
| M3 | 0.0004 | **4.1235** | **8.1662** | – | 1.0537 | 0.2817 |
| M5 | 2.6309 | 0.0187 | 1.3709 | – | 1.9668 | 1.6749 |

**TABLE 8.** Result of the ANOVA test (*F*-values) for quality scores obtained by the method proposed.

| | Babble speech noise | | | Street noise | | |
|---|---|---|---|---|---|---|
| SNR | –5 dB | 0 dB | 5 dB | –5 dB | 0 dB | 5 dB |
| M2 | 0.0001 | 0.0675 | 0.0000 | 0.0431 | 0.0231 | 0.0039 |
| M3 | 0.0047 | 0.0047 | 0.0008 | 0.0199 | 0.0070 | 0.0012 |
| M5 | 0.0013 | 0.0015 | 0.0002 | 0.0001 | 0.0008 | 0.0003 |

**TABLE 9.** Result of the ANOVA test (*F*-values) for subjective quality scores.

| | Babble speech noise | | | Street noise | | |
|---|---|---|---|---|---|---|
| SNR | –5 dB | 0 dB | 5 dB | –5 dB | 0 dB | 5 dB |
| M2 | 0.0002 | 0.1077 | 1.4978 | 0.67742 | 0.53348 | 0.40766 |
| M3 | 0.0607 | 0.5462 | **4.1795** | 0.00311 | 1.21187 | 0.6159 |
| M5 | 3.6195 | **8.8708** | **17.8488** | 0.70643 | 0.23492 | 0.69137 |

similar to a natural Lombard speech sample. This occurred as some outliers in the listeners' responses. Also, it transpired that glottal vocoder Lombard speech samples were clearly preferred in terms of naturalness. Both of the carried out tests differ from those performed by us, thus a direct comparison is not possible. It seems, however, that the speech samples may be more easy to evaluate in the pairwise listening test than in MUSHRA. Therefore, this type of subjective tests will be utilized in our future studies [59].

### D. STATISTICAL RESULTS

In order to check whether differences between the measurements are statistically significant, a statistical analysis of the results was performed. For this purpose, the one-way ANOVA test was employed, which we used to measure the variation between the utterances with the Lombard effect (LS) and their models. The null hypothesis ($H_0$) states that the utterance and its model are from populations with the same means. The decision rule to reject this hypothesis can be expressed by the following formula:

$$reject\ H_0 \quad if\ F > F_{critical}(1-\alpha) \tag{11}$$

where $F$ is the calculated test statistic, and $F_{critical}$ is the critical value taken from the $F$-distribution table. Details on how to perform analyses using ANOVA as well as critical values of $F$-distribution, are given in the textbook of Tabachnick and Fidell [60]. The test significance level $\alpha$ equals 0.05 (based on the ITU-R Recommendation BS.1534-1 [2]).

The test results are given in Tables 7-9. Differences that are statistically significant are highlighted in bold font.

## VI. CONCLUSION AND FURTHER INVESTIGATIONS

Based on the subjective test results, it can be observed that in most cases, the sinusoidal model with phase preserving, and both source-filter models hold a leading position when it comes to mimicking natural Lombard speech features. However, the ANOVA analysis shows that only the differences between the source-filter model with waveform-based parameter and natural Lombard speech are statistically significant. In contrast, the statistical test shows that the difference between the source-filter model with the aperiodicity parameter and natural speech is not statistically significant. Hence, the superiority of source-filter models over other models utilized is proved. This model may serve as a basis for Lombard speech modeling. In a future study, various modifications to this model will be tested to improve speech quality in adverse noise conditions.

The results of the objective test defined by the ITU-T P.563 recommendation and the method based on acoustic parameters, proposed by the authors, differ to some extent. Following this, a listening test should be performed to check whether the proposed measure overestimates the quality of the models, or the MOS-LQO underestimates it. Based on the results we obtained, it may be observed that the proposed measure correlates more closely with the results of the subjective evaluation than the outcomes from the ITU-T P.563 recommendation. Moreover, when analyzing the results of the subjective test, it was found that when babble speech was applied, the MOS values at a SNR of −20 dB, −15 dB, or −10 dB were not reliable. Meanwhile, our proposed method was stable in terms of signal quality in the presence of this noise. These facts support the assumption that the measure proposed is a good predictor of the Lombard effect, and it can be utilized as an indicator of speech quality.

The experiment results show that the SNR level threshold for which a particular model stopped working was −5 dB in the case of babble speech for both objective quality evaluation techniques. For street noise distortions, the thresholds were 5 dB and 0 dB, respectively, for the P.563 indicator and the method based on signal parameters. Meanwhile, in the case of subjective evaluation, the threshold was 0 dB in the case of street noise, and for babble speech noise, it was not possible to determine based on the experiment carried out. Even though the assumptions and conditions of our study differ from other research works, we can conclude that the outcomes of our research in the context of the threshold at which the model stops working are better compared to the state-of-the-art. As already said, results obtained by Michelsanti *et al.* [57] refer to two types of subjective tests, namely MUSHRA and speech intelligibility test. For the AO-L at −5 dB SNR, the result was approx. 25 points, whereas for 5 dB SNR returned approx. 50 points. Depending on the distortion type and speech models tested, we have got approx. 60 points in the MUSHRA test for the babble speech distortion set to −5 dB, and for the 5 dB SNR case between approx. 76 and 54 points. For the street noise distortion, the results were, however, much lower, i.e., approx. 30 points for −5 dB, and 60 points for 5 dB SNR.

With regard to subjective test scores, the results are comparable to some extent to those achieved by Michelsanti *et al.* [57]. Estimated averaged MOS-LQO values for babble speech and street noise distortions at 5 dB

SNR (a case when recordings containing only sentences were used in the evaluation process), depends to some extent on the speech model used. So, for the source-filter model with the aperiodicity parameter (M2) at the −20 dB SNR we have got approx. 2.4 and for 5 dB approx. 3.4. This is an average of the results obtained for all speech models in the case of speech babble distortion. However, for the source-filter model with a waveform-based parameter (M3) and babble speech distortion, the highest value we have got was 4.06. On the other hand, estimated averaged quality scores for babble speech and street noise distortions obtained by the method based on acoustical parameters derived from speech were even higher, i.e., 4.69 for babble speech distortion and 4.83 for the street distortion. In both cases, this occurred for the source-filter model with the aperiodicity parameter (M2). In comparison, Michelsanti *et al.* [57] reported averaged scores of PESQ between 1.346 (for −20 to 5 dB) to 3.127 (for 10 to −30 dB) for the audio-only case with the Lombard effect (AO-L). In contrast, in our case we have got values higher than 4.5 at 20 dB SNR. An important conclusion may also be derived from the state-of-the-art; namely, it seems that the converted speech samples may be easier for the subjects to be evaluated in the pairwise listening test than in MUSHRA.

Overall, even though the processing and synthesis of Lombard speech have made many advances over recent years, there still exists a need to improve speech synthesis models to make them more robust in adverse SNR conditions. There are many areas of applications awaiting such a feature, i.e., hearing aid algorithms, speech enhancement, language understanding in noisy environments, and automatic conversion in Text-to-Speech, to name a few. Based on the investigations performed, two directions of study development are foreseen by the authors. The first one, already mentioned, is to apply modifications to the best speech synthesis model in order to make it more robust over many SNRs, and the second one concerns proposing a quality measure that is better suited for speech in noisy environments than the measures contained in the standard. As ITU-T Recommendation P.563 was initially created for predicting the subjective quality of telephony applications, thus this needs to be addressed.

## APPENDIX

The list of sentences:

1) Wykonuj polecenia organów Straży Pożarnej i Policji!
2) Kieruj się w stronę wyjścia ewakuacyjnego!
3) Proszę jak najszybciej opuścić budynek!
4) Zakaz korzystania z wind!
5) Proszę wezwać ochronę!
6) Czy wśród nas jest lekarz?
7) Gdzie znajduje się najbliższe wyjście ewakuacyjne?
8) Gdzie znajduje się sprzęt gaśniczy?
9) Czy ktoś potrafi udzielić pierwszej pomocy?
10) Czy została wezwana karetka pogotowia.
11) Nie ma zagrożenia, to nie jest pożar.
12) W prawym skrzydle budynku zostało wyłączone zasilanie.
13) Wszystkie pomieszczenia zostały przeszukane.
14) Winda uległa awarii, proszę poruszać się schodami.
15) Za chwilę nastąpi ewakuacja wszystkich osób z budynku

The list of words:

1) korytarz
2) alarm
3) gaśnica
4) ewakuacja
5) wypadek
6) ochrona
7) lekarz
8) pojedynczo
9) zatrzymaj się
10) biegnij

## REFERENCES

[1] *Single-Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications*, document Rec. ITU-T P.563, 2004. [Online]. Available: https://www.itu.int/rec/T-REC-P.563-200405-I/en
[2] *Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*, document Rec. ITU-R BS.1534-1, International Telecommunications Union, 2003. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1534-1-200301-S/en
[3] T. Biberger, J.-H. Fleßner, R. Huber, and S. Ewert, "An objective audio quality measure based on power and envelope power cues," *J. Audio Eng. Soc.*, vol. 66, nos. 7–8, pp. 578–593, Aug. 2018.
[4] C. Valentini-Botinhao, J. Yamagishi, and S. King, "Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise?" in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 1837–1840.
[5] G. Korvel, O. Kurasova, and B. Kostek, "An attempt to create speech synthesis model that retains lombard effect characteristics," in *Proc. 16th Int. Joint Conf. e-Bus. Telecommun. (ICETE)*, 2019, pp. 286–295.
[6] E. Lombard, "Le signe de l'élévation de la voix (translated from French)," *Ann. des Mal. l'oreille du larynx*, vol. 37, no. 2, pp. 101–119, 1911.
[7] H. Boril, P. Fousek, and H. Hőge, "Two-stage system for robust neutral/Lombard speech recognition," in *Proc. 8th Annu. Conf. Int. Speech Commun. Assoc.*, 2007, pp. 1074–1077.
[8] M. Garnier, N. Henrich, and D. Dubois, "Influence of sound immersion and communicative interaction on the lombard effect," *J. Speech, Lang., Hearing Res.*, vol. 53, no. 3, pp. 588–608, Jun. 2010.
[9] G. Bapineedu, "Analysis of Lombard effect speech and its application in speaker verification for imposter detection," Ph.D. dissertation, Lang. Technol. Res. Centre, Int. Inst. Inf. Technol., Bengaluru, India, 2010.
[10] S. Seshadri, L. Juvela, O. Räsänen, and P. Alku, "Vocal effort based speaking style conversion using vocoder features and parallel learning," *IEEE Access*, vol. 7, pp. 17230–17246, 2019.
[11] A. K. Ho, J. L. Bradshaw, R. Iansek, and R. Alfredson, "Speech volume regulation in Parkinson's disease: Effects of implicit cues and explicit instructions," *Neuropsychologia*, vol. 37, no. 13, pp. 1453–1460, Dec. 1999.
[12] Y. Lu and M. Cooke, "Speech production modifications produced by competing talkers, babble, and stationary noise," *J. Acoust. Soc. Amer.*, vol. 124, no. 5, pp. 3261–3275, Nov. 2008.
[13] R. Patel and K. W. Schell, "The influence of linguistic content on the Lombard effect," *J. Speech, Lang., Hearing Res.*, vol. 51, no. 209, pp. 209–5101, Feb. 2008.
[14] H. L. Pick, G. M. Siegel, P. W. Fox, S. R. Garber, and J. K. Kearney, "Inhibiting the lombard effect," *J. Acoust. Soc. Amer.*, vol. 85, no. 2, pp. 894–900, Feb. 1989.
[15] S. A. Zollinger and H. Brumm, "The lombard effect," *Current Biol.*, vol. 21, no. 16, pp. R614–R615, Aug. 2011.
[16] A. S. Therrien, J. Lyons, and R. Balasubramaniam, "Sensory attenuation of self-produced feedback: The Lombard effect revisited," *PLoS ONE*, vol. 7, no. 11, pp. 1–7, 2012.

[17] S. G. Adams and A. E. Lang, "Can the lombard effect be used to improve low voice intensity in Parkinson's disease?" *Int. J. Lang. Commun. Disorders*, vol. 27, no. 2, pp. 121–127, Jan. 1992.

[18] E. T. Stathopoulos, J. E. Huber, K. Richardson, J. Kamphaus, D. DeCicco, M. Darling, K. Fulcher, and J. E. Sussman, "Increased vocal intensity due to the Lombard effect in speakers with Parkinson's disease: Simultaneous laryngeal and respiratory strategies," *J. Commun. Disorders*, vol. 48, pp. 1–17, Mar. 2014.

[19] R. Marxer, J. Barker, N. Alghamdi, and S. Maddock, "The impact of the Lombard effect on audio and visual speech recognition systems," *Speech Commun.*, vol. 100, pp. 58–68, Jun. 2018, doi: 10.1016/j.specom.2018.04.006.

[20] E. Jokinen, M. Takanen, M. Vainio, and P. Alku, "An adaptive post-filtering method producing an artificial Lombard-like effect for intelligibility enhancement of narrowband telephone speech," *Comput. Speech Lang.*, vol. 28, no. 2, pp. 619–628, Mar. 2014.

[21] J. Blauert, "Assessing the 'quality-of-the-acoustics' at large," *J. Audio Eng. Soc.*, vol. 67, nos. 1–2, pp. 5–12, 2019.

[22] D. Z. Rodríguez, R. L. Rosa, F. L. Almeida, G. Mittag, and S. Moller, "Speech quality assessment in wireless communications with MIMO systems using a parametric model," *IEEE Access*, vol. 7, pp. 35719–35730, 2019.

[23] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.

[24] G. P. Kafentzis, O. Rosec, and Y. Stylianou, "Robust full-band adaptive sinusoidal analysis and synthesis of speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 6260–6264.

[25] K. K. Paliwal and L. Alsteris, "Usefulness of phase in speech processing," in *Proc. IPSJ Spoken Lang. Process. Workshop*, Gifu, Japan, 2003, pp. 1–6.

[26] J. Deng, X. Xu, Z. Zhang, S. Fruhholz, and B. Schuller, "Exploitation of phase-based features for whispered speech emotion recognition," *IEEE Access*, vol. 4, pp. 4299–4309, 2016.

[27] T. Koc and T. Ciloglu, "Nonlinear interactive source-filter models for speech," *Comput. Speech Lang.*, vol. 36, pp. 365–394, Mar. 2016.

[28] J. W. Beauchamp, "Comparison of vocal and violin vibrato with relationship to the source/filter model," in *Studies in Musical Acoustics and Psychoacoustics*, no. 36. Cham, Switzerland: Springer, 2017, pp. 201–221.

[29] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: Vocoder revisited," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 2, Apr. 1997, pp. 1303–1306.

[30] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. Syst.*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[31] G. Degottex, P. Lanchantin, and M. Gales, "A log domain pulse model for parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 1, pp. 57–70, Jan. 2018.

[32] L. Li, Y. Nankaku, and K. Tokuda, "A Bayesian approach to voice conversion based on GMMs using multiple model structures," in *Proc. Interspeech*, Florence, Italy, 2011, pp. 661–664.

[33] A. Suni, T. Raitio, M. Vainio, and P. Alku, "The GlottHMM speech synthesis entry for Blizzard Challenge 2010," in *Proc. Blizzard Challenge Workshop*, 2010, pp. 1–6.

[34] G. Korvel, V. Šimonytė, and V. Slivinskas, "A phoneme harmonic generator," *Inf. Technol. Control*, vol. 45, no. 1, pp. 7–12, Mar. 2016.

[35] G. Pyž, V. Šimonytė, and V. Slivinskas, "Developing models of lithuanian speech vowels and semivowels," *Informatica*, vol. 25, no. 1, pp. 55–72, Jan. 2014.

[36] H. Kawahara, A. D. Cheveigné, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, 2005, pp. 537–540.

[37] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoust. Sci. Technol.*, vol. 27, no. 6, pp. 349–353, 2006.

[38] M. Morise, "PLATINUM: A method to extract excitation signals for voice synthesis system," *Acoust. Sci. Technol.*, vol. 33, no. 2, pp. 123–125, 2012.

[39] M. Morise, H. Kawahara, and H. Katayose, "Fast and reliable F0 estimation method based on the period extraction of vocal fold vibration of singing voice and speech," in *Proc. Audio Eng. Soc. Conf., 35th Int. Conf., Audio Games Audio Eng. Soc.*, Feb. 2009, pp. 77–81.

[40] M. Morise, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Commun.*, vol. 67, pp. 1–7, Mar. 2015.

[41] D. P. W. Ellis. (2003). *Sinewave and Sinusoid+Noise Analysis/Synthesis in MATLAB Online Web Resource*. Accessed: Nov. 2019. [Online]. Available: http://www.ee.columbia.edu/dpwe/resources/matlab/sinemodel/

[42] T. Abe, T. Kobayashi, and S. Imai, "The IF spectrogram: A new spectral representation," in *Proc. ASVA*, 1997, pp. 423–430.

[43] D. P. W. Ellis and R. J. Weiss, "Model-based monaural source separation using a vector-quantized phase-vocoder representation," in *Proc. IEEE Int. Conf. Acoust. Speed Signal Process. Proc.*, vol. 5, May 2006, pp. 1–4.

[44] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 3, pp. 323–332, May 1999.

[45] T. Abe, T. Kobayashi, and S. Imai, "Harmonics tracking and pitch extraction based on instantaneous frequency," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 1995, pp. 756–759.

[46] *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, document Rec. P.862, Feb. [Online]. Available: http://www.itu.int/rec/T-REC-P.862/en ITU-T

[47] *Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs*, document ITU-T Rec. P.862, International Telecommunications Union, 2001.

[48] J. G. Beerends, A. P. Hekstra, A. W. Rix, and M. P. Hollier, "Perceptual evaluation of speech quality (PESQ): The new ITU standard for end-to-end speech quality assessment part II—Psychoacoustic model," *J. Audio Eng. Soc.*, vol. 50, pp. 765–778, Oct. 2002.

[49] R. K. Dubey and A. Kumar, "Comparison of subjective and objective speech quality assessment for different degradation/noise conditions," in *Proc. Int. Conf. Signal Process. Commun. (ICSC)*, Mar. 2015, pp. 261–266.

[50] I. Kraljevski, S. Chungurski, I. Stojanovic, and S. Arsenovski, "Synthesized speech quality evaluation using ITU-T P.563," in *Proc. 18th Telecommun. Forum TELFOR*, 2010, pp. 590–593.

[51] G. Korvel, A. Kurowski, B. Kostek, and A. Czyzewski, "Speech analytics based on machine learning," in *Machine Learning Paradigms* (Intelligent Systems Reference Library), vol. 149, G. Tsihrintzis, D. Sotiropoulos, L. Jain, Eds. Cham, Switzerland: Springer, 2019, pp. 129–157.

[52] B. Kostek, A. Kupryjanow, P. Zwan, W. Jiang, Z. W. Raś, M. Wojnarski, and J. Swietlicka, "Report of the ISMIS 2011 contest: Music information retrieval," in *Proc. Int. Symp. Methodol. Intell. Syst.*, Berlin, Germany, 2011, pp. 715–724.

[53] A. Rosner, B. Schuller, and B. Kostek, "Classification of music genres based on music separation into harmonic and drum components," *Arch. Acoust.*, vol. 39, no. 4, pp. 629–638, Mar. 2015.

[54] *Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*, document Rec. ITU-R BS.1534-2, International Telecommunications Union, 2014. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1534-2-201406-S/en

[55] S. Zieliński, "On some biases encountered in modern audio quality listening tests (part 2): Selected graphical examples and discussion," *J. Audio Eng. Soc.*, vol. 64, nos. 1–2, pp. 55–74, Feb. 2016.

[56] *webMUSHRA*. Accessed: Nov. 2019. [Online]. Available: https://www.ee.columbia.edu/~pwe/resources/matlab/sinemodel/

[57] D. Michelsanti, Z.-H. Tan, S. Sigurdsson, and J. Jensen, "Deep-learning-based audio-visual speech enhancement in presence of lombard effect," *Speech Commun.*, vol. 115, pp. 38–50, Dec. 2019.

[58] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[59] A. R. López, S. Seshadri, L. Juvela, O. Räsänen, and P. Alku, "Speaking style conversion from normal to lombard speech using a glottal vocoder and Bayesian GMMs," in *Proc. Interspeech*, Aug. 2017, pp. 1363–1367.

[60] B. G. Tabachnick and L. S. Fidell, *Experimental Designs Using ANOVA*. Belmont, CA, USA: Thomson/Brooks/Cole, 2007.

**GRAŽINA KORVEL** received the B.S. degree in mathematics and the M.S. degree in informatics from Vilnius Pedagogical University (recently Vytautas Magnus University Education Academy), Lithuania, in 2007 and 2009, respectively, and the Ph.D. degree from the Institute of Data Science and Digital Technologies, Vilnius University, in 2013. She is currently a Senior Researcher with the Institute of Data Science and Digital Technologies. Her research interests include speech signal processing, developing of mathematical models, applications of soft computing, and computational intelligence. The main scientific results have been published in 27 papers and discussed at 40 national and international conferences. Some of her works received the Diploma for the Best Presentation. She is a three-time Winner of the Lithuanian Academy of Sciences Young Scientist Award. She received acknowledgment of the Prime Minister of Lithuania for her obtained scientific results, in 2013 and 2019. She is a Reviewer of many scientific journals, a member of regional and international scientific societies, and has been the Session Organizer in international conferences.

**KRZYSZTOF KĄKOL** received the M.S. degree in sound engineering from the Gdańsk University of Technology, in 2001. He has been working for many years as a Software Engineer, a System Analyst, and DevOps and Solution Architect. He is currently employed at PGS Software S.A., polish software house, where he is working as the Solutions Architect and the Manager. His commercial and research interests include data pipelines, data analysis and processing, and data science, especially connected with neural networks.

**OLGA KURASOVA** received the Ph.D. degree in computer science from the Institute of Mathematics and Informatics, Vytautas Magnus University, Lithuania, in 2005. She is currently employed as a Principal Researcher and a Professor at the Institute of Data Science and Digital Technologies, Vilnius University. Her research interests include data mining methods, optimization theory and applications, artificial intelligence, neural networks, visualization of multidimensional data, multiple criteria decision support, parallel computing, and image processing. She is the author of more than 70 scientific publications.

**BOŻENA KOSTEK** (Senior Member, IEEE) is currently a Professor with the Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology (GUT), Poland. She has presented more than 600 scientific papers in journals and international conferences. She has also led a number of research projects. She has supervised more than 200 master's theses and 16 Ph.D. theses. She also published three books related to multimedia applications. Her main scientific research interests include psychoacoustics, music information retrieval, multimedia, cognitive and behavioral processing, and applications of machine learning methods to the mentioned domains. She is a Corresponding Member of the Polish Academy of Sciences and a Fellow of the Audio Engineering Society (AES). She was a recipient of many prestigious awards for research, including those of the Prime Minister of Poland (twice), Ministry of Science, and the Polish Academy of Sciences. From 2003 to 2007, she was appointed to serve as the AES Vice President for the Central Europe Region, from 2003 to 2007, and the AES Governor, from 2007 to 2009. She was also elected as the AES Vice President for the Central Europe Region, from 2009 to 2011. She is the Editor-In-Chief of the *Journal of the Audio Engineering Society*.