Problems of Forensic Sciences 2013, vol. 95, 642–650

ISSN 1230-7483

NOISE REDUCTION IN AUDIO RECORDINGS FOR FORENSIC PURPOSES USING AN ALGORITHM FOR SHORT-TIME SPECTRAL ANALYSIS

Marcin MICHAŁEK

Institute of Forensic Research, Kraków, Poland

Summary

Forensic speech to text transcriptions from evidence recordings, which are usually intensively distorted, are carried out in almost every forensic audio analysis. This is connected with a need to correct recordings to improve the quality of the speech signal (forensic audio enhancement). The purpose of this work was to develop our own software to reduce various kinds of noises contained in evidence audio recordings, enabling improvement of speech signal intelligibility. During project implementation, an algorithm for adaptive noise reduction was used with short-time spectral amplitude estimation (MMSE STSA). Analyses conducted using the constructed software allowed us to evaluate the effectiveness of the reconstruction of the speech signal for different kinds of noises and *SNR* values, and define the threshold below which correct transcription is difficult or impossible. The results made it possible to indicate what kind of noise has the most destructive effect on the speech signal and what conditions must be fulfilled by recordings in order to enable a correct transcription. The impact of the applied algorithm on possible distortions of the speech signal after reconstruction was also estimated. The designed and validated software for noise reduction will be a useful tool that allows more effective audio analysis in forensic practice.

Key words

Forensic audio analysis; Audio recording; Noise cancellation; Spectral analysis.

Received 8 May 2013; accepted 24 July 2013

1. Introduction

In evidence audio recordings, as practice shows, very often there are numerous noises characterized by a high volume. These noises may result from the way of recording, acoustic conditions during the event, and the recording technique that was used [1, 2]. Basic forensic audio analysis consists in auditory analysis and transcription of utterances registered in a recording. [1, 3, 10, 12]. High volume and frequency characteristics of noise very often effectively mask the speech signal, which limits the feasibility of listening tests and transcription [1]. It is for this reason that correction of recordings is so important for forensic audio analysis. The aim is to improve the quality and intelligibility of the recorded speech [1, 3, 10]. In addition to the classical signal filtering methods, adaptive noise reduction methods play an important part in such cases. They are based on algorithms that dynamically adapt the parameters of filtering to the current noise, often varying over time [4, 6, 7, 11]. Algorithms for analysis and estimation of the speech signal in the frequency domain, such as minimum mean-square error short-time spectral amplitude estimation, are particularly noteworthy [4, 5]. This algorithm allows for effective reduction of the noise in the recordings and signal-to-noise ratio (*SNR*) improvement with low computational complexity and low distortion. The great advantage of this algorithm is that it can be used to reduce noise in single-channel systems, which is often a feature of analyzed evidence recordings.

2. Materials and methods

2.1. The research material

At the outset of this research, appropriate audio recordings were collected, containing both recorded speech and noises. The developed database contains recordings with utterances in Polish collected from seventeen people, i.e. five women and twelve men. The recordings contain three kinds of noise: inside a running car, traffic and impulse noise. These noises are characterized by the following time-frequency characteristics:

– inside a car: long-term, narrow-band, multiple harmonics;
– traffic: long-term, broadband;
– impulse: short-term, broadband.

The rationale for the selection of such representative noises was to ensure diversity of time-frequency characteristics and their occurrence in evidence recordings. The parameters of these noises also allow us to assess the potential of their reduction by using the constructed software, which is the subject of further research.

All of the collected recordings were made using a portable digital recorder with the following parameters: recording format using MP3 compression, 192 kbps, 44,100 Hz, monaural mode. All recordings with utterances were divided up into representative excerpts lasting a few seconds, and for further research, two excerpts were used from each speaker. For the test recordings prepared in this way, recorded noises were added, assuming *SNR* values in the range of –28 to 0 dB, as shown in Table I.

TABLE I. *SNR* VALUES ASSUMED FOR TEST RECORDINGS CONTAINING SPEECH SIGNAL AND ADDED NOISES

Kind of noise	Signal-to-noise ratio (*SNR*) [dB]					
Inside a car	–6	–12	–15	–18	–24	–28
Traffic	–3	–6	–9	–12	–18	–21
Impulse	0	–3	–4	–6	–9	–12

SNR values were chosen experimentally for each type of noise to determine the threshold values between comprehensible and incomprehensible utterances, after noise reduction. Based on the assumption of signals additivity [4], using MATLAB software, an algorithm was developed to automatically sum each test recording with noises, in order to obtain *SNR* values. For this purpose, the following formula was used:

$$SNR = 10\log_{10}\frac{S_{power}}{N_{power}}, \quad \{1\}$$

where: S_{power} denotes the power of the useful signal (speech signal), N_{power} denotes the power of the noise.

When designing algorithms and software to reduce noise in audio recordings, the following major assumption was made: such software should be used primarily for recordings such as evidential ones, which are often of low quality. In order to approximate such parameters, in addition to recording by using compression, the sampling frequency of test recordings was lowered to 11,025 Hz. The research material thus consisted of source recordings containing speech and noise, recordings after adding noise and recordings after its reduction by using the constructed software. A total of 1226 audio files were collected for detailed analysis.

2.2. Applied methods

In evidence audio recordings, there are noises with different time and frequency characteristics that may change over time. In the main software module, amplitude estimation of the reconstructed speech signal is based on minimum mean-square error short-time spectral analysis (minimum mean square error short time spectral amplitude estimator, MMSE STSA estimator) [4, 5]. It is assumed that the noisy speech signal $y(t)$ can be represented by the equation:

$$y(t) = x(t) + d(t), \quad \{2\}$$

where $x(t)$ is the clean speech signal and $d(t)$ is an additive noise.

Let $X_n(k) = |X_n(k)| \exp(j\alpha_k)$,

$Y_n(k) = |Y_n(k)| \exp(j\theta_k)$

and let $D_n(k)$ denote discrete Fourier transforms of the clean and noisy speech signal and the additive noise, respectively, n be the number of the analyzed frame and k be the index in the frequency domain. Bearing the above in mind, it is possible to calculate the *a priori SNR* value on the basis of the formula:

$$\varepsilon_n(k) = \frac{E\{|X_n(k)|^2\}}{E\{|D_n(k)|^2\}}, \quad \{3\}$$

and *a posteriori SNR* by:

$$\gamma_n(k) = \frac{|Y_n(k)|^2}{E\{|D_n(k)|^2\}}, \quad \{4\}$$

where $E\{\cdot\}$ denotes the expected value.

Because of the fact that $E\{|D_n(k)|^2\}$ isn't available, $\gamma_n(k)$ is approximated by $\hat{\gamma}_n(k)$, which can be written as follows:

$$\hat{\gamma}_n(k) = \frac{|Y_n(k)|^2}{\lambda_n(k)}, \qquad \{5\}$$

where $\lambda_n(k)$ is the power spectrum of the noise signal. It can be calculated by averaging the power spectrum from frames which include only noise, i.e. without a speech signal. Taking into consideration $\hat{\gamma}_n(k)$ in the current frame and $\hat{\gamma}_{n-1}(k)$ in the previous frame, *a priori SNR* $\varepsilon_n(k)$ is approximated by $\hat{\varepsilon}_n(k)$ which is calculated in a decision-directed approach by:

$$\hat{\varepsilon}_n(k) = \alpha\, \hat{\gamma}_{n-1}(k)\mathrm{G}^2_{\mathrm{n-1}}(\mathrm{k}) + (1-\alpha)P[\,\hat{\gamma}_n(k) - 1], \qquad \{6\}$$

where $P(\cdot)$ is the rectifying function which ensures a positive value of the estimator. The α coefficient assumes values in the range of $0 < \alpha < 1$ and allows selection of the most optimal *a priori SNR*. $G_n(k)$ is the spectral gain which can be calculated with *a priori SNR* and *a posteriori SNR*. Estimated spectral amplitude of the clean speech signal can be calculated by multiplying $G_n(k)$ by the correlation function $R_n(k)$.

The presented algorithm allows for an effective increase in the ratio of useful signal to noise. The whole recording, which is subjected to noise reduction, is divided into smaller pieces (frames) with a length of 16 ms. The above calculations of amplitude estimation are performed for each frame with 50% overlapping, while the parameters of the noise reduction algorithm are adaptively adjusted in response to a change in noise. The α coefficient was determined experimentally and was taken as 0.96, but it can be modified. Taking into account the non-linear nature of sound perception by the human ear, the algorithm used in the frequency domain uses a logarithmic scale.

An additional module of the designed software uses algorithms for classical low-pass or high-pass filtering. It is used optionally for pre-processing and preparing the recording for the main noise reduction process. The band-pass filtering module uses an algorithm to design a stable Butterworth IIR filter (infinite impulse response) for specified parameters, i.e. the order and cut-off frequencies, and an algorithm for filtering the recording with the calculated filter [8, 9, 13, 14]. The stop-band of such a filter should not overlap with the useful signal band, i.e. speech.

Software which operates in the MATLAB computing environment was designed on the basis of the above described algorithms. It enables presentation of the time and frequency characteristics of the recording and allows the recording to be listened to and saved before, during and after noise reduction. Recordings are saved as audio files in the standard WAVE format. A block diagram containing all the modules in the developed computer program is shown in Figure 1.

3. Research results

The collected test audio recordings, containing speech signals and added noises with various *SNR*s, were processed to reduce the noises. The developed software together with the MMSE STSA module was used for this purpose.

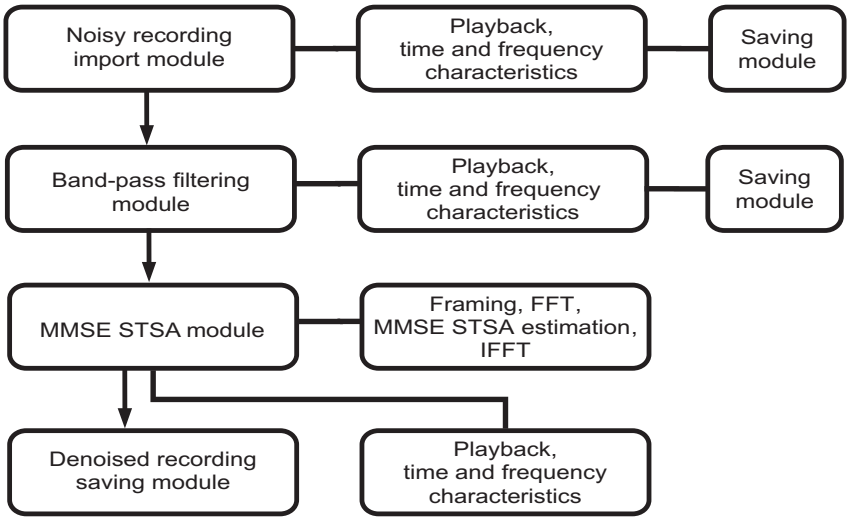

Fig. 1. A block diagram of the developed software to reduce noise in audio recordings.

As part of the validation and evaluation of the effectiveness of the noise reduction method used, the following analyses were conducted:

– comparison of the amplitude estimation algorithm used in the frequency domain with similar methods, such as Wiener and Kalman filtering and spectral subtraction;
– auditory analysis and assessment of speech intelligibility in test recordings after reduction of three kinds of noises and with different values of *SNR*;
– determining *SNR* threshold values in dB below which correct speech to text transcription of recordings after noise reduction is difficult or impossible;
– investigating the effect of various noises on the possibility of speech signal reconstruction;
– introducing potential distortions of speech by the used method, and comparing with source recordings.

When determining *SNR* thresh old values defining the threshold of speech intelligibility, the following rating scale was applied:

– rating 2: all utterances intelligible or some of them with high probability;
– rating 1: utterances intelligible with high probability or some unintelligible;
– rating 0: most or all utterances unintelligible.

For each test recording between a rating of 2 and 1, the *SNR* threshold value above which it is possible to make correct audio analysis and transcription after noise reduction, but below which correct transcription is difficult was determined. In turn, noise with an *SNR* less than the threshold value between the 1 and 0 rating does not allow a correct transcription to be made. The range of applicability of the method and its limitations were determined in this way. Tests were performed independently by four experts in the field of forensic audio analysis.

As regards the collected test recordings, the research showed that methods using adaptive Wiener and Kalman filtering and spectrum subtraction have good noise reduction and improvement in *SNR*. However, Wiener and Kalman filters have a high computational complexity compared to the MMSE STSA algorithm. In turn, the method of spectral subtraction generates so called musical noise, which is a negative phenomenon, especially in the case of a speech signal of low *SNR*. The amplitude estimation algorithm that was used has the most optimal noise reduction and *SNR* improvement, at low cost computing, slight distortion and, furthermore, it takes into account the characteristics of human sound perception.

Test recordings after noise reduction were subjected to auditory analysis and assessment of intelligibility of the recorded utterances, using the adopted grading scale. Figures 2 and 3 shows the results of this analysis.

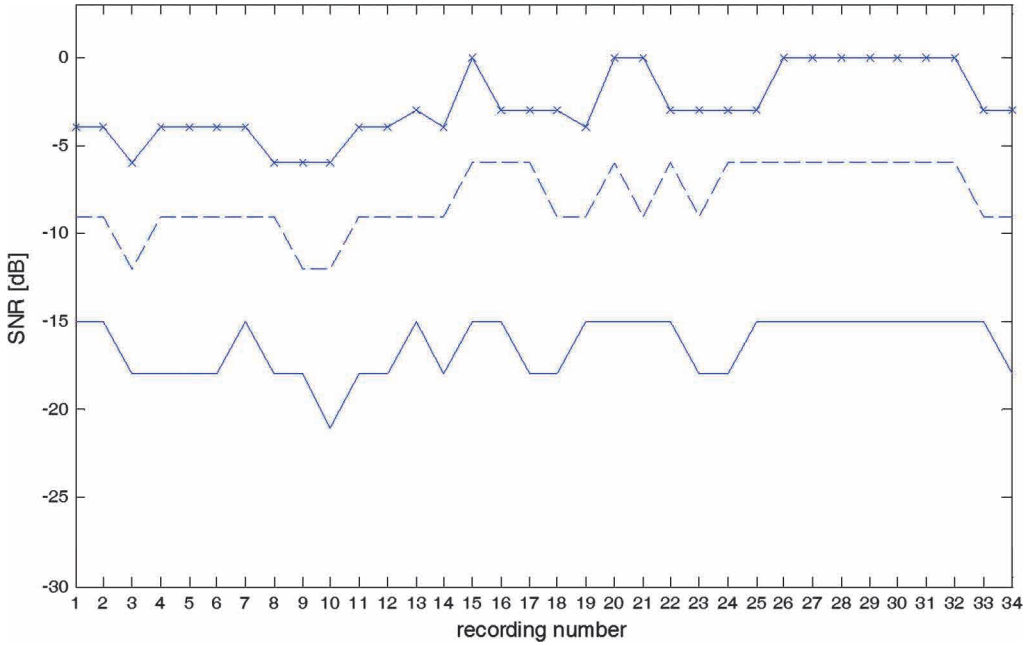

Fig. 2. Determined *SNR* threshold values between 2 and 1 rating for recordings after noise reduction: inside a car (solid line), traffic (line – –) and impulse (line ×–×).

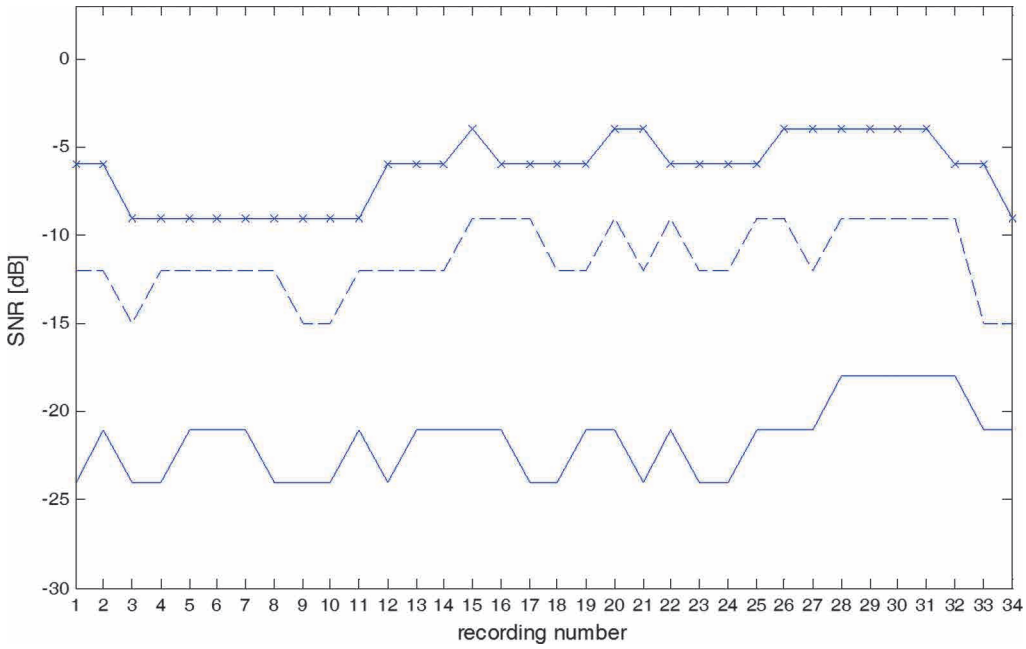

Fig. 3. Determined *SNR* threshold values between 1 and 0 rating for recordings after noise reduction: inside a car (solid line), traffic (line – –) and impulse (line ×–×).

Average *SNR* threshold values were calculated for the analyzed test recordings on the basis of the determined *SNR* threshold values, as shown in Table II.

TABLE II. AVERAGE *SNR* THRESHOLD VALUES BETWEEN SPEECH INTELLIGIBILITY RATINGS IN RECORDINGS AFTER NOISE REDUCTION

	Noise		
	Inside a car	Traffic	Impulse
Average *SNR* threshold value between 2 and 1 rating [dB]	–16.4	–8	–2.7
Standard deviation	1.69	1.91	2.02
Average *SNR* threshold value between 1 and 0 rating [dB]	–21.6	–11.4	–6.4
Standard deviation	2.06	2.06	1.92

The results of the assessment of speech intelligibility obtained for the test recordings allow us to conclude that it is possible to make correct speech to text transcriptions after noise reduction by the proposed method, if the *SNR* amounts to, on average, more than –16.4 (noise inside a car), –8 (traffic) and –2.7 dB (impulse). In turn, *SNR* values equal to or less than, on average, –21.6, –11.4 and –6.4 dB do not allow for correct transcription of utterances from the analysed test recordings (Table II).

The results indicate that impulse noises have the most destructive influence on the speech signal, traffic has less, and noise recorded inside a car has the least destructive effect. It was also found that impulse noises are the most bothersome when listening to a recording and making a transcription. Auditory analysis showed that the amplitude estimation algorithm used in the frequency domain introduced slight distortions of the speech signal, but they were observed mainly for voiceless vowels. However, they did not significantly affect the correctness of transcription of the analysed test recordings.

4. Conclusions

Under this project, software for reduction of noise in audio recordings and for improving speech intelligibility was designed and created. The study provided an opportunity to test the software and evaluate its features and usefulness for reducing noise in low quality recordings. The obtained results allowed us to conclude that the constructed software is a helpful tool in forensic audio analysis, both for correcting the

quality of audio recordings and making speech to text transcriptions. It will allow faster and more extensive auditory analysis. The algorithm used is characterized by optimal noise reduction with negligible distortions and, very importantly, by the possibility of its application to single-channel recordings. The developed software for noise reduction has been implemented and is used in current forensic practice in the Section of Speech and Audio Analysis of the Institute of Forensic Research.

References

1. Błasikiewicz S., Metoda odsłuchu szeptu i mowy intensywnie zakłóconej, *Problemy Kryminalistyki* 1971, 90, 159–183.
2. Błasikiewicz S., Bednarczyk W., Podstawowe zagadnienia kryminalistycznej identyfikacji osób na podstawie sygnału mowy za pomocą EMC, *Problemy Kryminalistyki* 1979, 142, 713–728.
3. Błasikiewicz S., Miściuk A., Wójcik W., Podstawowy zakres badań fonoskopijnych prowadzonych w Zakładzie Kryminalistyki KG MO, *Problemy Kryminalistyki* 1967, 67/68, 303–327.
4. Ephraim Y., Malah D., Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator, *IEEE Transactions on Acoustic*, *Speech and Signal Processing* 1984, 32, 1109–1121.
5. Ephraim Y., Malah D., Speech enhancement using a minimum mean-square error log-spectral amplitude estimator, *IEEE Transactions on Acoustic, Speech and Signal Processing* 1985, 33, 443–445.
6. Kustra G., Algorytmy identyfikacji i adaptacji w jednokanałowych systemach aktywnej redukcji hałasu, Materiały IX Sympozjum Naukowego „Nowości w technice audio i wideo", Wydawnictwo SIGMA-NOT, Warszawa 2002.
7. Kustra G., Zastosowanie adaptacyjnych sieci neuronowych do aktywnej redukcji hałasu, Materiały IX Sympozjum Naukowego „Nowości w technice audio i wideo", Wydawnictwo SIGMA-NOT, Warszawa 2002.
8. Lyons R. G., Wprowadzenie do cyfrowego przetwarzania sygnałów, Wydawnictwa Komunikacji i Łączności, Warszawa 2000.
9. MATLAB User's Guide R2012b [http://www.mathworks.com/help/pdf_doc/allpdf.html, 2012].
10. Rzeszotarski J., Kompendium badań fonoskopijnych, *Prokuratura i Prawo* 2007, 7/8, 181–187.
11. Smoliński R., Cyfrowe systemy rekonstrukcji dźwięku, Materiały VII Sympozjum Naukowego „Nowości w technice audio", Warszawa 2000.
12. Suszczewski W., Ekspertyza fonoskopijna, [in:] Ekspertyza sądowa. Zagadnienia wybrane, Wójcikiewicz J. [ed.], Oficyna a Wolters Kluwer Business, Warszawa 2007.
13. Zalewski A., Cegieła R., MATLAB – obliczenia numeryczne i ich zastosowania, Wydawnictwo NAKOM, Poznań 1996.
14. Zieliński T. P., Cyfrowe przetwarzanie sygnałów. Od teorii do zastosowań, Wydawnictwa Komunikacji i Łączności, Warszawa 2007.

Corresponding author
Dr inż. Marcin Michałek
Instytut Ekspertyz Sądowych
ul. Westerplatte 9
31-033 Kraków
e-mail:mmichalek@ies.gov.pl

Problems of Forensic Sciences 2013, vol. 95, 642–650

REDUKCJA ZAKŁÓCEŃ W NAGRANIACH DŹWIĘKOWYCH NA POTRZEBY BADAŃ FONOSKOPIJNYCH Z ZASTOSOWANIEM ALGORYTMU DO KRÓTKOOKRESOWEJ ANALIZY SPEKTRALNEJ

1. Wstęp

W dowodowych nagraniach dźwiękowych, jak wykazuje praktyka, bardzo często występują liczne zakłócenia charakteryzujące się znacznym poziomem głośności. Zakłócenia te mogą wynikać ze sposobu nagrywania, warunków akustycznych podczas trwania zdarzenia oraz z użytej techniki rejestracji [1, 2]. Podstawowym badaniem fonoskopijnym jest analiza audytywna i spisanie treści wypowiedzi utrwalonych w nagraniu [1, 3, 10, 12]. Wysoki poziom głośności oraz charakterystyki częstotliwościowe zakłóceń niejednokrotnie skutecznie maskują sygnał mowy, co ogranicza możliwości wykonania badań odsłuchowych i spisania treści [1]. Z tego względu istotna dla badań fonoskopijnych jest korekcja nagrania, mająca za zadanie poprawę jakości i zrozumiałości utrwalonych wypowiedzi [1, 3, 10]. Oprócz klasycznych metod filtracji sygnałów, dużą rolę pełnią w tym przypadku adaptacyjne metody redukcji zakłóceń. Bazują one na algorytmach, które dynamicznie dostosowują parametry filtracji do bieżących zakłóceń, często zmiennych w czasie [4, 6, 7, 11]. Na szczególną uwagę zasługują algorytmy do analizy i estymacji sygnału mowy w dziedzinie częstotliwości, takie jak estymacja amplitudy w krótkookresowej analizie spektralnej z minimalizacją błędu średniokwadratowego [4, 5]. Algorytm ten pozwala na efektywną redukcję zakłóceń w nagraniach i poprawę stosunku sygnału do szumu (ang. signal-to-noise ratio, *SNR*) przy niewielkiej złożoności obliczeniowej i małych zniekształceniach. Dużą zaletą tego algorytmu jest to, że może on być wykorzystywany do redukcji zakłóceń w systemach jednokanałowych, co jest częstą cechą analizowanych nagrań dowodowych.

2. Materiał i metody

2.1. Materiał do badań

Na wstępie realizacji niniejszych badań zgromadzono odpowiednie nagrania dźwiękowe zarówno z utrwaloną mową, jak i z zakłóceniami. Stworzona baza zawierała nagrania wypowiedzi w języku polskim pobranych od siedemnastu osób, tj. pięciu kobiet i dwunastu mężczyzn. Zarejestrowano także nagrania zawierające trzy rodzaje zakłóceń: we wnętrzu samochodu z włączonym silnikiem, wynikające z ruchu ulicznego oraz impulsowe. Zakłócenia te odznaczały się następującymi charakterystykami czasowo-częstotliwościowymi:

- wnętrze samochodu: długookresowe, wąskopasmowe, liczne harmoniczne;
- ruch uliczny: długookresowe, szerokopasmowe;
- impulsowe: krótkookresowe, szerokopasmowe.

Uzasadnieniem wyboru takich reprezentatywnych zakłóceń było zapewnienie różnorodności charakterystyk czasowo-częstotliwościowych oraz ich występowanie w nagraniach dowodowych. Parametry tych zakłóceń pozwalają również na ocenę możliwości ich redukcji za pomocą skonstruowanego oprogramowania, co jest przedmiotem dalszych badań.

Wszystkie zgromadzone nagrania zarejestrowano za pomocą przenośnego rejestratora cyfrowego z następującymi parametrami: format zapisu wykorzystujący kompresję MP3, 192 kbps, 44 100 Hz, tryb monofoniczny. Całe nagrania z wypowiedziami podzielono na reprezentatywne, kilkusekundowe fragmenty i do dalszych badań wykorzystano po dwa fragmenty pochodzące od każdego mówcy. Do tak przygotowanych nagrań testowych dodano zarejestrowane zakłócenia, przyjmując wartości *SNR* w przedziale od –28 do 0 dB, co przedstawiono w tabeli I.

Wartości *SNR* dobrano eksperymentalnie dla każdego rodzaju zakłócenia, aby wyznaczyć, po redukcji zakłóceń, wartości graniczne pomiędzy wypowiedziami zrozumiałymi a niezrozumiałymi. Wykorzystując założenie addytywności sygnałów dźwiękowych [4], za pomocą oprogramowania MATLAB opracowano algorytm do automatycznego sumowania każdego nagrania testowego z zakłóceniami, aby uzyskać przyjęte wartości *SNR*. W tym celu posłużono się wzorem:

$$SNR = 10\log_{10}\frac{S_{power}}{N_{power}}\,, \qquad \{1\}$$

gdzie: S_{power} to moc sygnału użytecznego (sygnału mowy), N_{power} to moc sygnału zakłócenia.

Projektując algorytmy i oprogramowanie do redukcji zakłóceń w nagraniach dźwiękowych poczyniono główne założenie: oprogramowanie to powinno mieć zastosowanie przede wszystkim do takich nagrań, jak dowodowe, które często odznaczają się niską jakością zapisu. W celu przybliżenia do takich parametrów, oprócz rejestracji z użyciem kompresji, częstotliwość próbkowania nagrań testowych obniżono do wartości 11 025 Hz. Materiał do badań stanowiły zatem nagrania źródłowe zawierające mowę i zakłócenia, nagrania po dodaniu zakłóceń

i nagrania po ich redukcji za pomocą skonstruowanego oprogramowania. Do szczegółowej analizy zgromadzono łącznie 1226 plików dźwiękowych.

2.2. Zastosowane metody

W dowodowych nagraniach dźwiękowych występują zakłócenia o różnych charakterystykach czasowych i częstotliwościowych, które mogą zmieniać się w czasie. W głównym module oprogramowania estymacja amplitudy zrekonstruowanego sygnału mowy bazuje na krótkookresowej analizie spektralnej z minimalizacją błędu średniokwadratowego (ang. minimum mean square error short time spectral amplitude estimator, MMSE STSA estimator) [4, 5]. Zakłada się, że zakłócony sygnał mowy $y(t)$ można przedstawić równaniem:

$$y(t) = x(t) + d(t), \qquad \{2\}$$

gdzie $x(t)$ to niezakłócony sygnał mowy, zaś $d(t)$ to zakłócenie addytywne.

Niech $X_n(k) = |X_n(k)| \exp(j\alpha_k)$,

$$Y_n(k) = |Y_n(k)| \exp(j\theta_k)$$

oraz $D_n(k)$ będą dyskretnymi transformatami Fourierowskimi odpowiednio niezakłóconego i zakłóconego sygnału mowy oraz zakłócenia, n to numer analizowanej ramki sygnału, zaś k to indeks w dziedzinie częstotliwości. Biorąc powyższe pod uwagę, można wyznaczyć wartość *a priori SNR* na podstawie wzoru:

$$\varepsilon_n(k) = \frac{E\{|X_n(k)|^2\}}{E\{|D_n(k)|^2\}}, \qquad \{3\}$$

oraz wartość *a posteriori SNR* według zależności:

$$\gamma_n(k) = \frac{|Y_n(k)|^2}{E\{|D_n(k)|^2\}}, \qquad \{4\}$$

gdzie $E\{\cdot\}$ oznacza wartość oczekiwaną.

Ponieważ $E\{|D_n(k)|^2\}$ nie jest dostępne, $\gamma_n(k)$ jest aproksymowane przez $\hat{\gamma}_n(k)$, co można zapisać jako:

$$\hat{\gamma}_n(k) = \frac{|Y_n(k)|^2}{\lambda_n(k)}, \qquad \{5\}$$

gdzie $\lambda_n(k)$ to spektrum mocy sygnału zakłócającego. Można go obliczyć w wyniku uśrednienia mocy spektrum w ramkach sygnału zawierających jedynie zakłócenie, tj. bez sygnału mowy. Biorąc pod uwagę $\hat{\gamma}_n(k)$ w bieżącej ramce sygnału i $\hat{\gamma}_{n-1}(k)$ w ramce poprzedniej, *a priori SNR* $\varepsilon_n(k)$ jest aproksymowane przez $\hat{\varepsilon}_n(k)$ zgodnie z algorytmem decyzyjno-kierunkowym opisanym jako:

$$\hat{\varepsilon}_n(k) = \alpha\, \hat{\gamma}_{n-1}(k) G^2_{n-1}(k) + (1-\alpha) P[\hat{\gamma}_n(k) - 1], \qquad \{6\}$$

gdzie $P(\cdot)$ to funkcja korygująca zapewniająca wartość dodatnią estymatora. Współczynnik α przyjmuje wartości w przedziale $0 < \alpha < 1$ i pozwala on na wybór najbardziej optymalnego *a priori SNR*. Wartość $G_n(k)$ to wzmocnienie, które może być obliczone za pomocą *a priori SNR* i *a posteriori SNR* [4, 5]. Estymacja amplitudy niezakłóconego sygnału mowy $x(t)$ wyznaczana jest przez przemnożenie wartości $G_n(k)$ z funkcją korelacji $R_n(k)$.

Zaprezentowany algorytm pozwala na efektywne zwiększenie wartości stosunku sygnału użytecznego do zakłócenia. Całe nagranie, które zostaje poddane procesowi redukcji zakłóceń, dzielone jest na mniejsze fragmenty (ramki) o długości 16 ms. Przedstawione powyżej obliczenia estymacji amplitudy wykonuje się dla każdej ramki z zakładką 50%, zaś parametry algorytmu do redukcji zakłóceń dostosowywane są adaptacyjnie w przypadku zmiany zakłócenia. Współczynnik α wyznaczono doświadczalnie i przyjęto wartość 0.96, przy czym może być ona modyfikowana. Uwzględniając nieliniową charakterystykę percepcji dźwięków przez ucho ludzkie zastosowany algorytm w dziedzinie częstotliwości wykorzystuje skalę logarytmiczną.

Dodatkowy moduł projektowanego oprogramowania stosuje algorytmy służące do klasycznej filtracji dolno- lub górnoprzepustowej. Jest on stosowany opcjonalnie w celu tzw. preprocessingu i przygotowania nagrania do właściwego procesu redukcji zakłóceń. Moduł filtracji pasmowej wykorzystuje algorytm służący do zaprojektowania stabilnego filtra IIR (ang. infinite impulse response) Butterwortha dla podanych parametrów, tj. rzędu i częstotliwości granicznych, oraz algorytm do filtracji nagrania obliczonym filtrem [8, 9, 13, 14]. Pasmo zaporowe takiego filtra nie powinno pokrywać się z pasmem sygnału użytecznego, tj. mowy.

Na podstawie opisanych wyżej algorytmów skonstruowano oprogramowanie pracujące w środowisku obliczeniowym MATLAB. Umożliwia ono prezentację charakterystyk czasowych i częstotliwościowych nagrania oraz jego odsłuch i zapis przed, w trakcie i po redukcji zakłóceń. Nagrania zapisywane są do plików dźwiękowych w standardowym formacie wave. Schemat blokowy zawierający zestawienie wszystkich modułów w opracowanym programie komputerowym przedstawiono na rycinie 1.

3. Wyniki badań

Zgromadzone testowe nagrania dźwiękowe zawierające sygnał mowy wraz z dodanymi zakłóceniami o różnym stosunku *SNR* poddano procesowi redukcji tych zakłóceń. Wykorzystano do tego celu opracowane oprogramowanie wraz z modułem MMSE STSA.

W ramach walidacji i oceny skuteczności redukcji zakłóceń zastosowaną metodą przeprowadzono następujące badania:

– porównanie zastosowanego algorytmu estymacji amplitudy w dziedzinie częstotliwości względem podob-

nych metod, tj. filtracji Wienera, Kalmana i odejmowania komponentów spektrum;
– audytywną analizę i ocenę zrozumiałości wypowiedzi w nagraniach testowych po redukcji trzech rodzajów zakłóceń i o różnych wartościach *SNR*;
– wyznaczenie granicznych wartości *SNR* w dB, poniżej których prawidłowe spisanie treści wypowiedzi z nagrań po redukcji zakłóceń jest utrudnione albo niemożliwe;
– zbadanie wpływu różnych zakłóceń na możliwości rekonstrukcji sygnału mowy;
– wprowadzanie przez zastosowaną metodę ewentualnych zniekształceń mowy wraz z porównaniem z nagraniami źródłowymi.

Podczas wyznaczania granicznych wartości *SNR* określających próg zrozumiałości mowy zastosowano następującą skalę ocen:
– ocena 2: zrozumiałe wszystkie wypowiedzi lub niektóre z dużym prawdopodobieństwem;
– ocena 1: wypowiedzi zrozumiałe z dużym prawdopodobieństwem lub niektóre niezrozumiałe;
– ocena 0: większość lub wszystkie wypowiedzi niezrozumiałe.

Dla każdego nagrania testowego pomiędzy oceną 2 a oceną 1 wyznaczono graniczną wartość *SNR*, powyżej której możliwy jest poprawny odsłuch i spisanie treści wypowiedzi po redukcji zakłóceń, natomiast poniżej tej wartości spisanie treści jest utrudnione. Z kolei zakłócenia o *SNR* mniejszym niż graniczna wartość pomiędzy oceną 1 a 0 nie umożliwiają poprawnego spisania treści wypowiedzi. Wyznaczono w ten sposób zakres stosowalności metody i jej ograniczenia. Testy zostały wykonane niezależnie przez czterech ekspertów z zakresu fonoskopii.

W odniesieniu do zgromadzonych nagrań testowych przeprowadzone badania wykazały, że metody wykorzystujące adaptacyjną filtrację Kalmana i Wienera oraz odejmowania komponentów spektrum odznaczają się dobrą redukcją zakłóceń oraz poprawą stosunku *SNR*. Jednakże filtry Kalmana i Wienera cechują się dużą złożonością obliczeniową w stosunku do algorytmu MMSE STSA. Z kolei metoda odejmowania komponentów spektrum sygnału generuje tzw. szum muzyczny, co jest zjawiskiem negatywnym, zwłaszcza w przypadku sygnału mowy o niewielkim stosunku *SNR*. Zastosowany algorytm estymacji amplitudy wykazuje najbardziej optymalną redukcję zakłóceń i poprawę stosunku *SNR* przy niewielkim nakładzie obliczeniowym, nieznacznych zniekształceniach oraz uwzględnieniu charakterystyki percepcji dźwięków przez człowieka.

Nagrania testowe po redukcji zakłóceń poddano audytywnej analizie i ocenie zrozumiałości utrwalonych w nich wypowiedzi, wykorzystując przyjętą skalę ocen. Na rycinach 2 i 3 przedstawiono wyniki tej analizy.

Dla analizowanych nagrań testowych i na podstawie wyznaczonych wartości granicznych *SNR* obliczono ich wartości średnie, co przedstawiono w tabeli II.

Otrzymane wyniki oceny zrozumiałości wypowiedzi dla nagrań testowych pozwalają na konkluzję, że możliwe jest poprawne spisanie treści wypowiedzi po redukcji zakłóceń proponowaną metodą, jeśli *SNR* będzie wynosił, średnio, więcej niż –16,4 dB (zakłócenia wewnątrz samochodu), –8 dB (ruch uliczny) i –2,7 dB (zakłócenia impulsowe). Z kolei wartości *SNR* równe lub mniejsze niż, średnio, –21,6, –11,4 i –6,4 dB nie pozwalają na poprawne spisanie treści z analizowanych nagrań testowych (tabela II).

Uzyskane rezultaty wskazują, że najbardziej destrukcyjny wpływ na sygnał mowy mają zakłócenia o charakterze impulsowym, mniejszy wprowadzają zakłócenia ruchu ulicznego, zaś najmniejszy spośród badanych – zakłócenia zarejestrowane wewnątrz samochodu. Ustalono również, że zakłócenia impulsowe są najbardziej uciążliwe podczas odsłuchu nagrania i spisywania treści wypowiedzi. Analiza audytywna wykazała, że zastosowany algorytm estymacji amplitudy w dziedzinie częstotliwości wprowadza nieznaczne zniekształcenia sygnału mowy, przy czym obserwowano je głównie dla głosek bezdźwięcznych. Nie wpływały one jednak znacząco na poprawność spisywania treści nagrań testowych.

4. Wnioski

W ramach prac nad niniejszym projektem opracowano i stworzono oprogramowanie umożliwiające redukcję zakłóceń utrwalonych w nagraniach dźwiękowych oraz poprawę zrozumiałości mowy. Przeprowadzone badania pozwoliły na przetestowanie tego oprogramowania i ocenę jego możliwości oraz przydatności do redukcji zakłóceń dla nagrań charakteryzujących się niską jakością. Uzyskane wyniki pozwoliły na stwierdzenie, że skonstruowane oprogramowanie jest pomocnym narzędziem podczas wykonywania badań fonoskopijnych zarówno do korekcji jakości nagrań dźwiękowych, jak i spisywania utrwalonych wypowiedzi. Pozwoli ono na szybsze i obszerniejsze przeprowadzanie badań odsłuchowych. Zastosowany algorytm odznacza się optymalną redukcją zakłóceń przy nieznacznych zniekształceniach oraz, co bardzo istotne, możliwością jego stosowania do nagrań jednokanałowych. Opracowane oprogramowanie do redukcji zakłóceń zostało wdrożone i jest wykorzystywane w bieżącej działalności opiniodawczej Pracowni Analizy Mowy i Nagrań Instytutu Ekspertyz Sądowych.