INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Difference in Perceived Speech Signal Quality Assessment Among Monolingual and Bilingual Teenage Students

Przemyslaw Falkowski-Gilski

Gdansk University of Technology, Faculty of Electronics, Telecommunications and Informatics, Gdansk, Poland

przemyslaw.falkowski@eti.pg.edu.pl

Abstract

The user perceived quality is a mixture of factors, including the background of an individual. The process of auditory perception is discussed in a wide variety of fields, ranging from engineering to medicine. Many studies examine the difference between musicians and non-musicians. Since musical training develops musical hearing and other various auditory capabilities, similar enhancements should be observable in case of bilingual people. This paper examines the difference in perceived speech signal quality between students from monolingual and bilingual classes. The subjective study was carried out on a group of 30 people, with 15 individuals in each class, aged 16-18 years old, considering three languages: English, German, and Polish. Results of this study may aid researchers as well as professionals active in the field of auditory perception, hearing loss related with ageing, and of course evaluation of networks and services.

Index Terms: coding, compression, signal processing, speech perception, speech recognition

1. Introduction

Over the last years, numerous researchers studied human abilities of acquiring and processing acoustic information from the environment. Furthermore, the process of auditory perception itself has been discussed in a wide variety of fields, from engineering to medicine.

Some focus on psychoacoustic studies concerning environmental sounds that cannot be labeled as speech and/or music [1]. Other discuss whether musicians and musically-educated people possess better abilities of recognizing sources of sound than non-musicians. It is expected that musical training develops not only musical hearing, but also enhances various auditory capabilities. They may include factors such as recognition and detection thresholds in the presence of background noise or quiet, affecting the recognition-detection threshold gap (RDTG) [2]. At signal levels close to detection threshold, not all acoustic signatures of sound are clearly audible. Only a limited set of perceptual cues can be used for sound source recognition, audio signal annotation, and music information retrieval (MIR) [3, 4]. Moreover, noise can affect the learning as well as concentration skills [5, 6].

One may therefore expect that listeners with high auditory skills would be able to obtain more information than those who possess only average auditory skills. In this context, similar advantages should develop in case of bilingual people, especially when it comes to interpreting, understanding, and assessing the quality of speech.

2. Related Work

Cochlear hearing loss, related with ageing and harmful working conditions, is linked with the impairment of the active, nonlinear mechanisms in the inner ear. This results in a decreased sensitivity to low-intensity stimuli, as well as supra-threshold deficits, including loss of frequency selectivity, loudness recruitment and impaired temporal processing [7].

In case of speech at a positive signal-to-noise ratio (SNR) in stationary noise, rapid gain fluctuations, introduced by fast-acting compression, can elevate noise segments in speech pauses [8]. This imposes modulations on the background at rates similar to speech, and reduces the long-term output SNR.

Further information on hearing losses, concerning a study carried out on normal-hearing (NH) and hearing-impaired (HI) listeners, is described in [9]. Whereas, a study concerning hearing-aid dynamic range compression (DRC) among HI listeners, is presented in [10]. Additional information on short-term hearing aid and auditory perception may be found in [11, 12].

Another study [13] demonstrated that musicians, compared to non-musicians, possessed enhanced abilities of understanding speech in the presence of noise. Further information concerning the perception of speech and music signals by musicians and non-musicians, as well as a description of previous investigation including state-of-the-art on musician hearing enhancement (MHE) and related topics, may be found in [14].

In [15] authors investigated how age of acquisition influenced the perception of second-language (English) speech among Mexican-Spanish-speaking listeners. The study involved individuals who learned fluent English before age 6 (early bilingual) or after age 14 (late bilingual), as well as American-English speakers (monolingual). Results showed significant benefits for monolinguals and early bilinguals, compared with late bilinguals.

Another investigation [16] focused on whether bilinguals have a deficit in speech perception for their second language, compared with monolingual speakers, under harsh listening conditions. According to obtained results, bilingual and trilingual individuals performed similarly to monolinguals in quiet conditions. However, their performance declined with increasing noise.

A group of authors in [17] investigated the effects of bilingualism, noise, and reverberation, on speech perception among listeners with normal hearing. The study involved a group of 15 monolingual American English speakers and 12 Spanish-English bilinguals, who learned English prior to 6 years of age (with no noticeable foreign accent).

http://dx.doi.org/10.21437/Interspeech.2021-16

Results showed poorer word recognition skills in case of the bilingual group, rather than monolinguals, under the presence of noise and reverberation.

The aforementioned studies inspired this one, concerning a group of teenagers from both monolingual and bilingual classes. However, this study is not focused on speech perception in harsh conditions, e.g. in the presence of noise, reverberation, etc. It focuses on how people with different backgrounds and language skills perceive speech signals. Moreover, whether or not monolinguals or bilinguals make a better group when it comes to evaluating systems and services, e.g. voice assistants.

3. About the Study

The aim of this study was to investigate the difference in perceived speech signal quality assessment among young people with no hearing disorders. This group consisted of teenage students, aged between 16-18 years old, including individuals from both bilingual and monolingual classes, with 15 people in each class.

3.1. Student background

The bilingual group of students consisted of Polish natives, who speak fluently in both Polish and English. Each individual started his or her public or private education in an English-speaking school. From primary up to secondary education, they attend all classes in English. Everyone uses both Polish and English in everyday life.

The monolingual group also consisted of Polish natives, who speak fluently in their mother tongue that is Polish. English is their second language of choice, in which they most often communicate abroad. In this case, each individual started his or her public or private education in a Polish-speaking school. From primary up to secondary education, they attend all classes in Polish, and use it in everyday life. Whereas English is the only subject taught in this language, which they attend on an advanced level.

Both groups of students also learn another language at school, namely German on an advanced level, during the same period. Due to this fact, it seemed interesting to study what are the differences when it comes to perceiving the quality of speech signal samples coded at different bitrates. It should be emphasized that none of them had hearing disorders.

3.2. Signal samples

The tested signal samples were sourced from ITU-T P.501 [18]. In this recommendation, the audio samples consist of sentences spoken in different languages. Each language offers two sentences spoken by two female and two male lectors. After a careful examination, samples from 4 language sets were selected, namely: American English (AE), British English (EN), German (GE), and Polish (PL).

The topic of digital audio signal processing and coding is well-described in [19, 20, 21]. Additional information on speech quality assessment may be found in [22, 23]. Whereas advancements in speech signal processing, including low bitrate and perceptual audio coding, are discussed in [24, 25, 26, 27, 28].

4. Quality Assessment

The main goal in any voice transmission system is to provide high-quality audio services at any time and everywhere. Furthermore bandwidth fluctuations, due to varying network conditions, may cause degradation in quality. This may be observed as end-to-end delay or error rate, described as technical quality of service (QoS), or subjective degradation of consumed content, viewed as user quality of experience (QoE).

Content and service providers aim at designing low-bitrate services, because the lower the bitrate, the more services are available to the end user. The topic of designing both subjective and objective quality assessment studies, covering a wide range of bitrates, is discussed in [29, 30].

4.1. Signal processing and coding

The original signal samples were available in the WAV 16-bit PCM format, mono audio mode. Next, each sample was coded using the Ogg Vorbis algorithm. The degraded signals samples were processed in 3 bitrates, namely: 8, 16, and 24 kbps. The sampling frequency was changed to 44.1 kHz, as in many popular terrestrial broadcasting and online streaming services.

This lossy coding scheme, investigated in [31, 32], was chosen due to the fact that, according to a preliminary questionnaire (see Fig. 1), the tested group of teenage students most often consumed audio content, consisting of speech and music signals, using Spotify (36%).

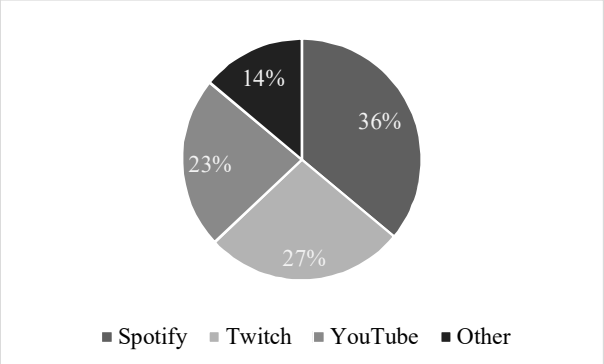

Figure 1: *Popular online streaming services.*

Among popular online streaming services, Twitch came in second (27%), YouTube came third (23%), whereas other (14%) included Apple Music, YouTube Music, Open.FM, etc. Further information on current trends in the consumption of multimedia content using online streaming platforms is available in [33].

4.2. Subjective evaluation

The subjective assessment was carried out using Beyerdynamic Custom One headphones in a 5-step mean opinion score (MOS) scale, with no reference signal available, ranging from 1 (bad quality) to 5 (excellent quality). The tests were carried out according to ITU-R BS 1284 [34]. A single session took approx. 10 minutes. Each person assessed the overall quality of speech signals individually after taking a training phase, in order to become familiar with the listening equipment as well as aim of the study. Participants were allowed to set the volume as desired during training phase.

5. Results

Results of this study, carried out on both monolingual and bilingual classes, including speech signal samples in American English (AE), British English (EN), German (GE), and Polish (PL), processed at 3 different bitrates, are shown in Figs. 2-4.

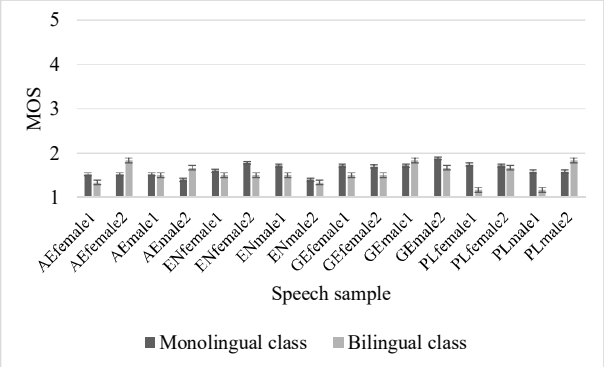

Figure 2: *Subjective quality evaluation of speech samples processed at 8 kbps.*

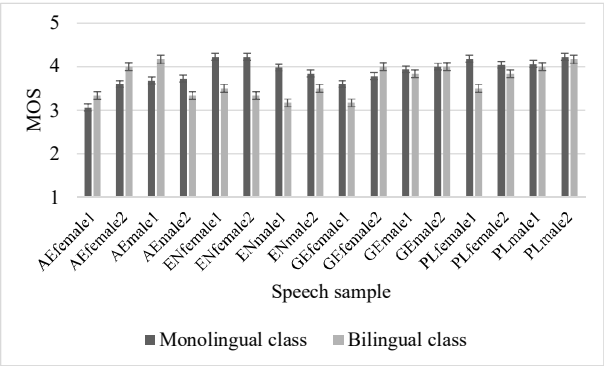

Figure 3: *Subjective quality evaluation of speech samples processed at 16 kbps.*

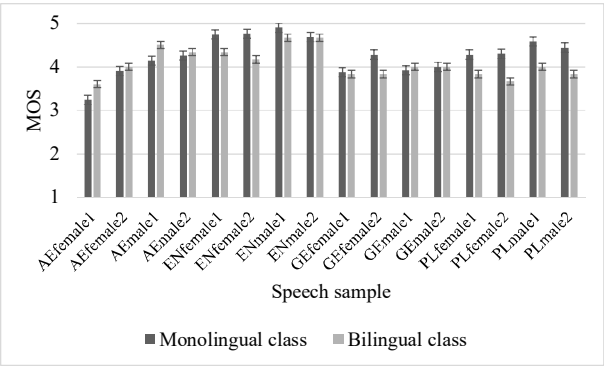

Figure 4: *Subjective quality evaluation of speech samples processed at 24 kbps.*

When statistically compiling obtained data using the analysis of variance (ANOVA) method, the confidence intervals were set to 95% (α=0.05). In all cases, the dispersion was less than 10% of the average values. It should be noted that neither individual was informed about the actual bitrate of the currently assessed signal sample. All stimuli were presented in a randomized way, and labeled as either male or female lector.

As shown, for the lowest bitrate of 8 kbps, according to both groups, the content quality was insufficient. All samples received an overall score of less than 2.0 (poor quality.) Furthermore, most often bilingual people were more critical and more sensitive to distortions caused by low-bitrate coding.

In case of the medium bitrate of 16 kbps, only a few samples (British English female as well as both male and female samples in Polish) were ranked as good (a score of approx. 4.0, but only in case of the monolingual class). Once again, the bilingual people proved to be more demanding users. A clear difference is visible in case of British English (EN) samples. Whereas for American English (AE) samples, results were quite the opposite. For the Polish (PL) language, obtained results were quite similar, with a slight difference in case of female samples. When it comes to German (GE), both classes ranked this set similarly.

In case of the highest bitrate of 24 kbps, bilingual individuals were more critical when it comes to evaluating speech samples in Polish (PL) and British English (EN). Results for German (GE) were quite similar, except for female lector 2. Whereas for American English (AE), in case of all lectors (both male and female), bilingual people ranked those samples evidently higher than monolinguals.

To sum up, generally speaking both groups ranked German (GE) speech samples similarly, as they learn this language during the same period of time. When it comes to the English dialect, monolingual people prefer British English (EN) rather than American English (AE), as they tend to rank this set of samples higher, regardless of the bitrate. On the other hand, bilingual people tend to give higher marks in case of the American English (AE) dialect, contrary to the monolinguals. This remark may be given regardless of the actual bitrate. Now in case of the Polish (PL) language, as both groups use it as their mother tongue (being Polish natives), bilingual people are more sensitive to distortions related with lossy compression and low-bitrate coding. In all cases, regardless of the processing bitrate, both male and female lector samples were ranked evidently lower, compared to monolingual individuals.

6. Discussion

In order to evaluate the impact of utilized bitrate (8, 16, and 24 kbps) on the perception of speech signal samples spoken in different languages (American English, British English, German, and Polish), among monolingual and bilingual students, a comparison has been performed. This set of results, shown in Figs. 5-8, is focused on particular spoken languages.

According to obtained results, when examining scores for the American English (AE) language set, it can be noticed that monolinguals are more demanding users, regardless of the actual bitrate. Obviously, as less experienced people when it comes to linguistics, they prefer the British dialect. On the other hand, in the remaining language sets, the situation is quite the opposite. Results for the British English (EN) set clearly show that bilingual people are more demanding users, regardless of the bitrate and type of lector.

When it comes to German (GE), both groups have less experience, compared to Polish and/or English language skills. Results for the evaluated classes are similar to another, yet bilinguals tend to be more critical most of the time.

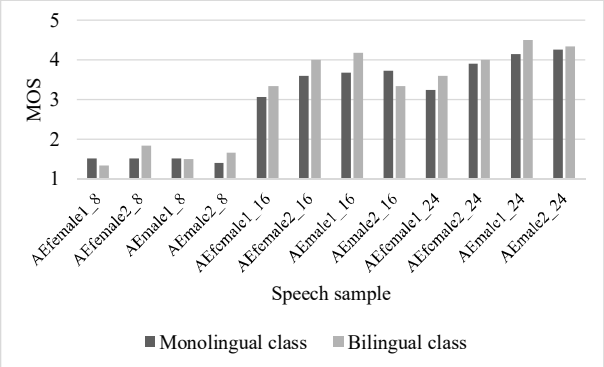

Figure 5: *Subjective quality evaluation of American English speech samples processed at different bitrates.*

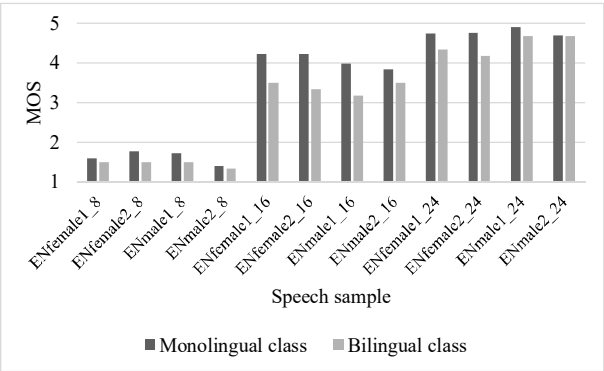

Figure 6: *Subjective quality evaluation of British English speech samples processed at different bitrates.*

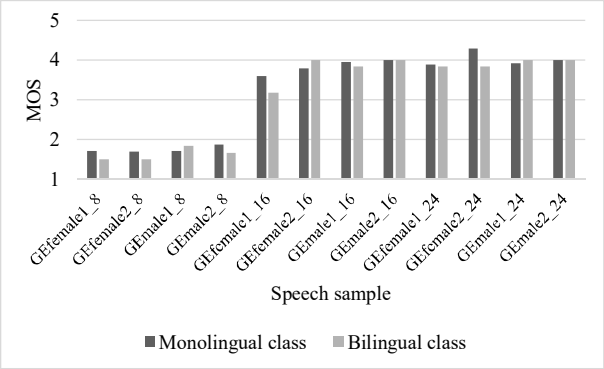

Figure 7: *Subjective quality evaluation of German speech samples processed at different bitrates.*

As Polish natives, both classes have similar language skills in their mother tongue. Despite this fact, once again, MOS scores of bilingual people tend to be lower compared to monolinguals.

To sum up, it can be concluded that language skills and human background determine individual linguistic preferences. It is clearly visible that monolingual people, who do not attend classes in both Polish and English, prefer the British English (EN) dialect over American English (AE). Furthermore, with the same experience in a foreign language, namely German (GE) learned at school during the same time period, both classes provided similar MOS scores.

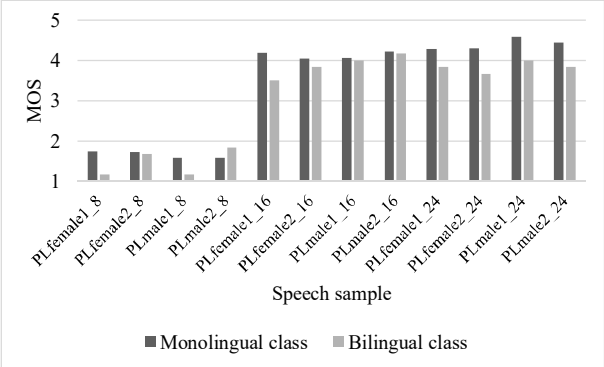

Figure 8: *Subjective quality evaluation of Polish speech samples processed at different bitrates.*

On the other hand, bilingual people obviously prefer the American English (AE) dialect. Moreover, although they learn and use their mother tongue Polish (PL) during the same time period as their monolingual colleagues, they are able to notice and point out more distortions present in the processed (coded) speech signal sample. It can be stated that superior language skills in a foreign language affect the perception of one's mother tongue.

7. Conclusions

As shown, language is based upon complex phonological systems. Speech communication involves multi-level cognitive processing systems that rely on detailed perceived characteristics of sound and on the discrimination of subtle changes of the acoustic signals. When listening to speech, the listener's attention is spread over a long time interval, which is required to extract the semantic message conveyed by sound. The difference in background of tested individuals may explain, at least partially, why the auditory skills developed by bilingual students resulted in visible lower ranking of audio content.

As observed, bilingual people are more demanding users. They are more sensitive to distortions related with lossy coding and processing of audio content. Under these circumstances almost every speech signal sample received a lower MOS score, compared to monolingual people. On the contrary, it can be said that monolinguals can appreciate quality content, as their subjective judgements were most often higher.

Moreover, when examining both groups of Polish natives, one can notice that when it comes to learning English, monolingual people prefer the British dialect (noticeably higher MOS scores), whereas bilinguals prefer the American dialect (noticeably higher MOS scores). This fact is clearly visible when examining their subjective judgements with respect to either British or American English. This remark is valuable not only for teachers from schools and universities, but also content creators and distributors, voice assistant test engineers, as well as other interested third parties.

Furthermore, it was shown that auditory skills related with foreign languages are strictly dependent on the learning time. As shown, the smallest difference in subjective judgements among both groups was observed in case of German. However, a small advantage may be observed in favor of the bilingual group. Similar remarks may be given in case of the mother tongue, namely Polish.

8. References

[1] B. Gygi, G. R. Kidd, and C. S. Watson, "Similarity and Categorization of Environmental Sounds," *Perception and Psychophysics*, vol. 69, pp. 839–855, 2007.

[2] K. Abouchacra, T. Letowski, and J. Gothie, "Detection and Recognition of Natural Sounds," *Archives of Acoustics*, vol. 32, no. 3, pp. 603–616, 2007.

[3] P. Fallgren, Z. Malisz, and J. Edlund, "How to Annotate 100 Hours in 45 Minutes," in *INTERSPEECH 2019 – 21th Annual Conference of the International Speech Communication Association, September 15-19, Graz, Austria, Proceedings*, 2019, pp. 341–345.

[4] B. Kostek, "Music Information Retrieval – The Impact of Technology, Crowdsourcing, Big Data, and the Cloud in Art," *Journal of the Acoustical Society of America*, vol. 146, no. 4, pp. 2946–2946, 2019.

[5] J. Kotus, M. Szczodrak, A. Czyżewski, and B. Kostek, "Long-term Comparative Evaluation of an Acoustic Climate in Selected Schools Before and After the Acoustic Treatment," *Archives of Acoustics*, vol. 35, no. 4, pp. 551–564, 2010.

[6] A. Zagubień and K. Wolniewicz, "The Assessment of Infrasound and Low Frequency Noise Impact on the Results of Learning in Primary School – Case Study," *Archives of Acoustics*, vol. 45, no. 1, pp. 93–102, 2020.

[7] B. C. J. Moore, *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues*. Chichester: John Wiley & Sons, 2007.

[8] G. Naylor and R. B. Johannesson, "Long-Term Signal-to-Noise Ratio at the Input and Output of Amplitude-Compression Systems," *Journal of the American Academy of Audiology*, vol. 20, no. 3, pp. 161–171, 2009.

[9] B. Kowalewski, T. May, M. Fręczkowski, J. Zaar, O. Strelcyk, E. N. MacDonald, and T. Dau, "Effects of Fast-Acting Hearing-Aid Compression on Audibility, Forward Masking and Speech Perception," in *Joint Conference – Acoustics 2018, September 11-14, Ustka, Poland, Proceedings*, 2018, pp. 1–6.

[10] E. Vilchuur, "Signal Processing to Improve Speech Intelligibility in Perceptive Deafness," *Journal of the Acoustical Society of America*, vol. 53, no. 6, pp. 1646–1657, 1973.

[11] T. Poremski, P. Szymański, and B. Kostek, "Assessment of the Effectiveness of a Short-term Hearing Aid Use in Patients with Different Degrees of Hearing Loss," *Archives of Acoustics*, vol. 44, no. 4, pp. 719–729, 2019.

[12] P. Szymański, T. Poremski, and B. Kostek, "The Influence of Time of Hearing Aid Use on Auditory Perception in Various Acoustic Situations," *Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1834–1835, 2018.

[13] A. Parbery-Clark, E. Skoe, C. Lam, and N. Kraus, "Musician Enhancement for Speech in Noise," *Ear and Hearing*, vol. 30, pp. 653–661, 2009.

[14] A. Miśkiewicz, T. Rościszewska, J. Żera, J. Majer, and B. Okoń-Makowska, "Detection and Recognition of Environmental Sounds by Musicians and Non-Musicians," *Archives of Acoustics*, vol. 43, no. 4, pp. 581–592, 2018.

[15] L. Hansberry Mayo, M. Florentine, and S. Buus, "Age of Second-Language Acquisition and Perception of Speech in Noise," *Journal of Speech, Language, and Hearing Research*, vol. 40, pp. 686–693, 1997.

[16] D. Tabri, K. M. Smith Abou Chacra, and T. Pring, "Speech Perception in Noise by Monolingual, Bilingual and Trilingual Listeners," *International Journal of Language and Communication Disorders*, vol. 46, no. 4, pp. 411–422, 2011.

[17] C. L. Rogers, J. J. Lister, D. M. Febo, J. M. Besing, and H. B. Abrams, "Effects of Bilingualism, Noise, and Reverberation on Speech Perception by Listeners with Normal Hearing," *Applied Psycholinguistics*, vol. 27, pp. 465–485, 2006.

[18] ITU Recommendation P.501, *Test Signals for Telecommunication Systems*, 2017.

[19] A. S. Spanias, "Speech Coding: A Tutorial Review," *Proceedings of the IEEE*, vol. 82, no. 10, pp. 1541–1582, 1994.

[20] T. Painter and A. Spanias, "Perceptual Coding of Digital Audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[21] A. Spanias, T. Painter, and V. Atti, *Audio Signal Processing and Coding*. Hoboken: John Wiley & Sons, 2007.

[22] S. Brachmański, "Objective Measure for Assessment of Speech Quality in Rooms," *Archives of Acoustics*, vol. 33, no. 4, suppl., pp. 177–182, 2008.

[23] K. Kondo, *Subjective Quality Measurement of Speech: Its Evaluation, Estimation and Applications*. Berlin-Heidelberg: Springer-Verlag, 2012.

[24] S. Kandadai and C. D. Creusere, "Scalable Audio Compression at Low Bitrates," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 969–979, 2008.

[25] J. Wang, P. Liu, J. Kong, and R. Ying, "Split Table Extension: A Low Complexity LVQ Extension Scheme in Low Bitrate Audio Coding," *IEEE Signal Processing Letters*, vol. 17, no. 1, pp. 59–62, 2010.

[26] Y. Yamamoto, T. Chinen, and M. Nishiguchi, "A New Bandwidth Extension Technology for MPEG Unified Speech and Audio Coding," in *ICASSP 2013 – IEEE International Conference on Acoustics, Speech and Signal Processing, May 26-31, Vancouver, Canada, Proceedings*, 2013, pp. 523–527.

[27] K. Seto and T. Ogunfunmi, "Packet-Loss Robust Scalable Speech Coding Using the Discrete Wavelet Transform," in *ISCAS 2014 – IEEE International Symposium on Circuits and Systems, June 1-5, Melbourne, Australia, Proceedings*, 2014, pp. 129–132.

[28] T. Vaillancourt, V. Malenovsky, R. Salami, Z. Liu, L. Miao, J. Gibbs, and M. Jelinek, "Advances in Low Bitrate Time-Frequency Coding," in *ICASSP 2015 – IEEE International Conference on Acoustics, Speech and Signal Processing, April 19-24, South Brisbane, Australia, Proceedings*, 2015, pp. 5913–5917.

[29] P. Gilski and J. Stefański, "Subjective and Objective Comparative Study of DAB+ Broadcast System," *Archives of Acoustics*, vol. 42, no. 1, pp. 3–11, 2017.

[30] P. Falkowski-Gilski, "Transmitting Alarm Information in DAB+ Broadcasting System," in *SPA 2018 – Signal Processing: Algorithms, Architectures, Arrangements, and Applications, September 19-21, Poznan, Poland, Proceedings*, 2018, pp. 217–222.

[31] H. Chen and T. L. Yu, "Comparison of Psychoacoustic Principles and Genetic Algorithms in Audio Compression," in *ICSEng 2005 – International Conference on Systems Engineering, August 16-18, Las Vegas, USA, Proceedings*, 2005, pp. 1–6.

[32] R. Korycki, "Detection of Montage in Lossy Compressed Digital Audio Recordings," *Archives of Acoustics*, vol. 39, no. 1, pp. 65–72, 2014.

[33] P. Falkowski-Gilski and T. Uhl, "Current Trends in Consumption of Multimedia Content Using Online Streaming Platforms: A User-Centric Survey," *Computer Science Review*, vol. 37, 100268, 2020.

[34] ITU Recommendation BS.1284, *General Methods for the Subjective Assessment of Sound Quality*, 2003.