INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2024, VOL. 70, NO. 3, PP. 597–602
Manuscript received June 17, 2024; revised July, 2024. doi: 10.24425/ijet.2024.149584

Lossy coding and bitrate effects on changes in formant frequencies in Japanese and English speech signals

Mateusz Andrzej Kucharski, and Stefan Brachmański

***Abstract*—Since speaker recognition and verification became heavily used technology, both in professional applications like forensics and more everyday ones, the question arose: what factors can impact results of those processes? One thing that may be important with respect to this subject is lossy coding, as some of the information contained in an original file is lost in the coding process. In the era of globalization, not only native languages or languages of neighboring countries are of interest to researchers, but also those quite far, especially from Asia – the biggest exporter of goods and services to Europe. Those economic relationships are usually connected with the interchange of personnel, which further shortens geographical distance. The article presents the results that are a continuation of research on the behavior of Japanese language formants. Earlier research focused on changes occurring for the first and second formants. This article presents changes observed for the third and fourth formants. The knowledge of these changes is indicated in the process of speaker identification in forensics using the spectrographic method. At the Department of Acoustics and Multimedia, Wroclaw University of Science and Technology and in many centers around the world, the auditory-spectrographic method is used, which is a combination of the aural and spectrographic methods. In the spectrographic part, a person is identified on the basis of a comparison of the formants' trajectory.**

***Keywords*—formants, formant frequency, bitrate, coding, lossy codecs, speech**

I. Introduction

THE goal of this work was to research lossy coding's impact on first four formant frequencies in speech with emphasis on third and fourth formants due to their decisiveness in speaker identification. Research concerning impact of coding techniques on formant changes in speech signals are conducted for many languages. Familiarity with this impact is very useful in forensic acoustics using spectral methods for speaker identification [1] [2] [3]. In this method, when making a decision, the expert analyzes the differences and similarities in the spectrograms of identical utterances from the evidence and comparative recording, comparing, among others, values and trajectories of the fundamental frequency of the laryngeal tone and formants (F1, F2, F3, F4) [4]. In every country, apart from native speakers, forensic expert might encounter criminals speaking other languages [5]. In Poland, languages most encountered in forensic research, apart from Polish, are Ukrainian, Russian and German, however Oriental languages, like Korean or Japanese, are appearing more and more often. Because of conducted, during forensic research, identification of person speaking Japanese language, a need to research impact of coding techniques on formant parameters changes for Japanese language arose. The work was split into two stages. First stage contained research of coding technique impact on pitch of the voice and first two formant frequencies [6], and the second stage – on third and fourth formant frequencies. In this article, the overall results of researching coding techniques impact on four first formant frequencies for Japanese speech signals. This article also expands the topic by examining the effect of different bitrates on formant frequency errors, as well as comparison of results for a foreign speaker and two native speakers. The research was using the same methodology as previously. The database was expanded using recordings of two native Japanese speakers provided with ITU-T P.501 recommendation and additional conversions of previously used files, this time using different bitrate values [7]. The research was also expanded by examining speech-focused codec Speex.

Authors are with Wroclaw University of Science and Technology, Wroclaw, Poland (e-mail: mateusz.kucharski@pwr.edu.pl, stefan.brachmanski@pwr.edu.pl).

II. Methodology

In this research, previously acquired, as well as additional data was used. Data was extracted from audio files of eight Japanese sentences included in ITU-T P.501 recommendation recorded in environment meeting the requirements of ITU-T P.800 by one of the authors of this article, who is not a Japanese native speaker, and converted into different lossy codec formats [8] [7] [9]. The codecs chosen were some of the most popular all-purpose codecs: MP3, AAC, OGG and WMA. There were two recordings of the same material made during two separate occasions using condenser microphone and open license software. They were also made using both 48 kHz and 16 kHz sampling frequency. The files are monophonic and were coded in 16 bit PCM. Additional data was acquired by converting in the same way audio files included in P.501 document. P.501 recommendation splits previously mentioned eight sentences between four speakers, two of which are female and two are male. Sentences 1 (Kare wa ayu wo tsuru

meijin desu.) and 2 (Kodai ejiputo de jusshinhō no genre ga tsukuraremashita.) are spoken by first female speaker, 3 (Dokusho no tanoshisa wo shitte kudasai.) and 4 (Ningen no kachi wa chishiki wo dō katsuyō suruka de kimarimasu.) by second female speaker, 5 (Kanojowo settoku shiyōtoshitemo mudadesu.) and 6 (Sono mukasi garasu wa taihen mezurashii monodeshita.) by first male speaker and 7 (Chikagoro no kodomotachi wa hiyowa desu.) and 8 (Igirisujin wa ameno nakawo heikide nurete arukimasu.) by second male speaker. Because the author is male, it was decided to only focus on examining sentences 5 to 8, because differences between male and female voices might have impact on results. Later, one of the recording sessions' material was also converted, also with 48 kHz and 16 kHz, using Speex codec, which is designed specifically for speech signals. Additionally, some of the previously used original audio files were converted into chosen codecs with different bitrate values. Firstly, the third and fourth formants from author's recordings were examined. Data regarding those formants was extracted during research of the first and second formant and was later compared in the same ways. OGG and WMA files were compared directly, while MP3 and AAC files were compared using syllabic division method. This was implemented due to differences in file lengths that were discovered after conversions. While OGG and WMA files had the same or very similar length to the original wave file, MP3 and AAC files were up to 85 ms longer. Thus, data extracted from OGG and WMA files was simply subtracted from its equivalent from the original files. As MP3 and AAC data could not be compared in this manner, sentences were split into syllables (in Japanese meaning – syllables representing Japanese kana characters), as accurately as possible and then data regarding formants was extracted [10]. Based on extracted data points containing information on formant frequencies mean values for each syllable was calculated and those values were compared between the original WAVE file and corresponding converted file. In both methods' cases absolute values were taken. Data extraction in both cases was done using Praat software with a default frequency resolution of 250 frequency steps. However, because Praat does not support all of the used codecs, the files had to be converted back to WAVE. The results of both previous research of F1 and F2 as well as current research of F3 and F4 are presented on Fig. 1-4 and seem to confirm hypothesis that the difference in formant frequencies between original and converted files tend to rise with the number of the formant. This is the case for all four codecs, however it is not always true in case of singular formant points, what can be seen for example with the first set of MP3 files, where the third formant has represented smaller differences than the second formant. Additionally, as vowel triangles which are a representation of position of vowel relative to each other are often based on formants, example of formant triangles presenting those differences for vowel "a" can be seen on Fig. 8-11. While these diagrams are used to represent vowels and as such use first two formant frequencies, they are also a good way to represent differences caused by lossy coding. As such it was proposed to utilize this representation method for the third and fourth formant frequencies as well to further demonstrate the errors caused by lossy coding.

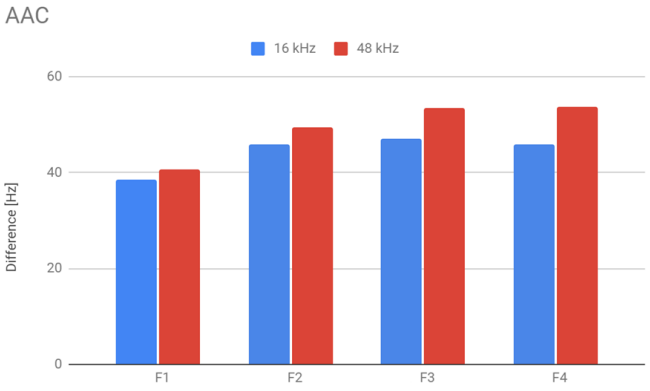

Fig. 1. Mean values of formant frequency differences introduced by AAC codec presented in Hz for 16 kHz and 48 kHz sampling frequencies

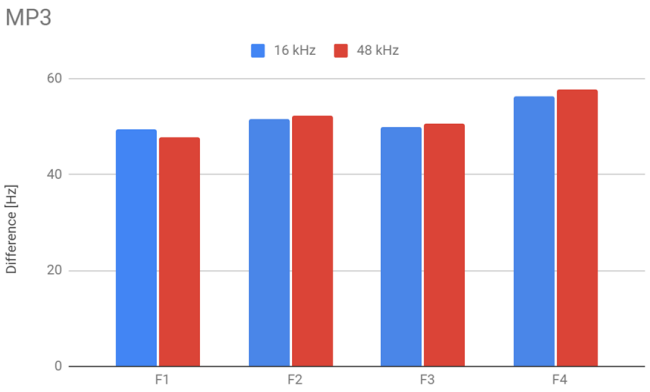

Fig. 2. Mean values of formant frequency differences introduced by MP3 codec presented in Hz for 16 kHz and 48 kHz sampling frequencies

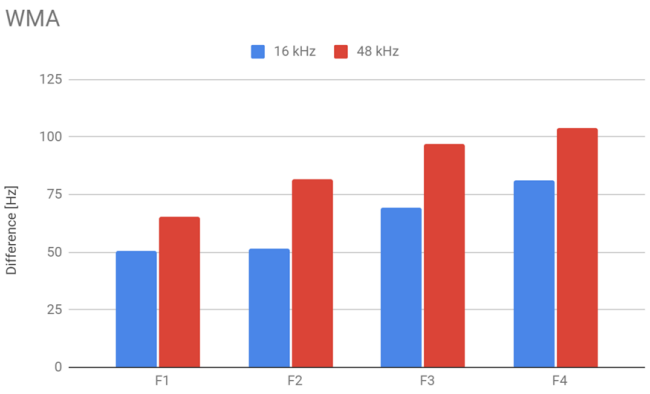

Fig. 3. Mean values of formant frequency differences introduced by WMA codec presented in Hz for 16 kHz and 48 kHz sampling frequencies

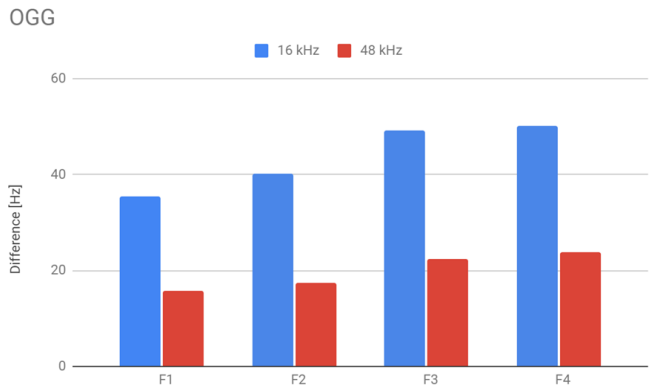

Fig. 4. Mean values of formant frequency differences introduced by OGG codec presented in Hz for 16 kHz and 48 kHz sampling frequencies

As the whole previous part of the research was done using recordings of only one speaker, the next step was to check if differences caused by lossy codecs have similar values for different speakers. For this purpose, recordings included in ITU-T P.501 recommendation were used. Only speakers of the same gender as the author were chosen, as differences between male and female voices could impact the results [11]. The results are presented in Table I-IV, which contain rounded off to the nearest integer mean differences for both male native speakers. Unfortunately, because the native speakers recorded only two sentences each, the amount of data is smaller than in case of authors' recordings, however, it can be seen that the results have similar values. Differences are the smallest for OGG codec (10 to 26 Hz difference), slightly bigger for MP3 and AAC (22 to 46 Hz for MP3 and 28 to 67 Hz for AAC) and the biggest for WMA (35 to 87 Hz).

TABLE I
MEAN VALUES OF FORMANT DIFFERENCES FOR NATIVE SPEAKER WITH AAC CODEC

AAC	$\Delta F1$ [Hz]	$\Delta F2$ [Hz]	$\Delta F3$ [Hz]	$\Delta F4$ [Hz]
Native speaker 1	34	28	29	56
Native speaker 2	31	36	56	67

TABLE II
MEAN VALUES OF FORMANT DIFFERENCES FOR NATIVE SPEAKER WITH MP3 CODEC

MP3	$\Delta F1$ [Hz]	$\Delta F2$ [Hz]	$\Delta F3$ [Hz]	$\Delta F4$ [Hz]
Native speaker 1	24	37	28	22
Native speaker 2	27	35	40	46

TABLE III
MEAN VALUES OF FORMANT DIFFERENCES FOR NATIVE SPEAKER WITH WMA CODEC

WMA	$\Delta F1$ [Hz]	$\Delta F2$ [Hz]	$\Delta F3$ [Hz]	$\Delta F4$ [Hz]
Native speaker 1	59	82	87	74
Native speaker 2	35	46	75	78

TABLE IV
MEAN VALUES OF FORMANT DIFFERENCES FOR NATIVE SPEAKER WITH OGG CODEC

OGG	$\Delta F1$ [Hz]	$\Delta F2$ [Hz]	$\Delta F3$ [Hz]	$\Delta F4$ [Hz]
Native speaker 1	12	21	26	20
Native speaker 2	10	12	15	11

Additionally it was decided to also check English speakers in similar fashion, however this time speaking English. For this purpose Google speech commands dataset was chosen, as it contains bigger variety of speakers. This dataset version contains 64 721 one second 16 kHz WAV files, containing 30 short English words spoken by variety of speakers. For the experiment dataset's first fifty speakers saying word "cat" were chosen [12]. From those files, as well as their encoded counterparts, formant data containing vowel "a" was extracted and compared. Results are presented on Fig. 5.

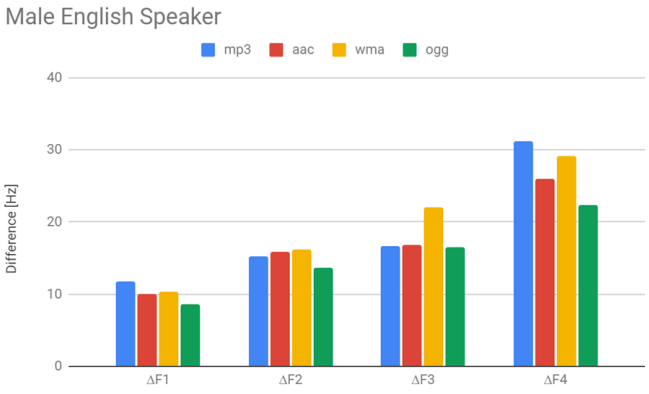

Fig. 5. Mean values of formant differences for male English speakers with four codecs

As the dataset contains a lot more individual speakers than ITU-Ts, it was also checked if there are any differences between genders. For this first fifty female speakers saying word "cat" were chosen and undergone the same procedure as their male counterparts. Results are presented on figure 6.

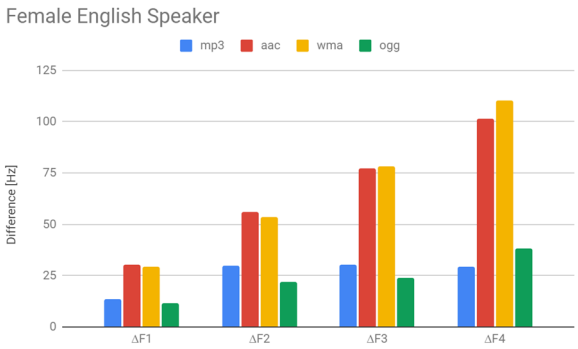

Fig. 6. Mean values of formant differences for female English speakers with four codecs

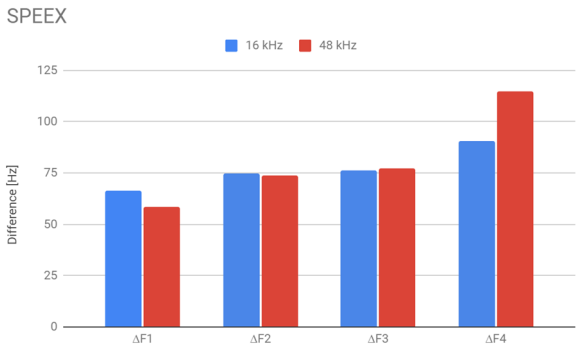

Fig. 7. Mean values of formant differences for Speex codec

As it can be seen on Fig. 5 the differences are also present, similar, or slightly lower than overall results for Japanese and Polish speakers, however it must be taken into consideration that only single vowel was examined in case of English speakers and it is probable that results for multiple syllables will be higher. In case of female English speakers MP3 and OGG codecs behave in similar fashion to male speakers, however for AAC and WMA noticeably higher values can be observed. The standard deviation of mean values for all speakers were calculated and while male English speakers display similar values, about 20-50 for all codecs [6], female English speakers display lower values, about 15-30 for all codecs. Because until now, only most popular, all-purpose codecs were examined, a question about speech-focused codecs arose. Therefore, the Speex codec was chosen to be examined. Data was mostly the same, except only the first recording sessions' material was used. It was also examined using two versions of sampling frequency: 16 kHz and 48 kHz. Data was extracted using syllabic division method, like it was done with MP3 and AAC files. The results are presented on Fig. 7. As it can be seen, the values of differences for all formants are far higher than what was achieved for all of the all-purpose codecs. All the values exceed 50 Hz difference and reach as high as 114 Hz for the fourth formant for files with 48 kHz sampling rate. Also interesting is that both the 16 and 48 kHz files share similar difference values, except for the fourth formant. However, it must be noted that previous research of all-purpose codecs was conducted on files with relatively high bitrate – 192 kbps, but Speex codec does not support bitrate values above 44 kbps [13].

Final aspect that was examined during this research was impact of the bitrate on differences between original and encoded files. Previously, bitrate value for encoding was usually set at 192 kHz, however it was suggested that lowering it might increase the differences due to lowered amount of information in a file. For this experiment, one of the original files, provided by ITU was used. It contained two sentences spoken by female speaker. It was converted into four previously examined codecs, AAC, MP3 WMA and OGG using four different bitrate values: 128, 96, 64 and 32 kbps. The only exception for this was WMA codec, which doesn't support 32 kbps [14]. Results are presented in Table V-VIII. The results for OGG and WMA codecs support previously mentioned thesis, as differences clearly increase in value as the bitrate lowers. The MP3 codec also exhibited slight increase, but this time with some fluctuations. However, the AAC codec seems mostly unaffected.

TABLE V
MEAN VALUES OF FORMANT DIFFERENCES FOR AAC CODEC WITH DIFFERENT BITRATE VALUES

AAC	$\Delta F1$ [Hz]	$\Delta F2$ [Hz]	$\Delta F3$ [Hz]	$\Delta F4$ [Hz]
32 kbps	40	70	71	58
64 kbps	42	65	65	58
96 kbps	34	63	47	46
128 kbps	39	68	58	45

TABLE VI
MEAN VALUES OF FORMANT DIFFERENCES FOR MP3 CODEC WITH DIFFERENT BITRATE VALUES

MP3	$\Delta F1$ [Hz]	$\Delta F2$ [Hz]	$\Delta F3$ [Hz]	$\Delta F4$ [Hz]
32 kbps	32	61	57	59
64 kbps	24	36	47	50
96 kbps	21	35	38	33
128 kbps	22	29	38	57

TABLE VII
MEAN VALUES OF FORMANT DIFFERENCES FOR WMA CODEC WITH DIFFERENT BITRATE VALUES

WMA	$\Delta F1$ [Hz]	$\Delta F2$ [Hz]	$\Delta F3$ [Hz]	$\Delta F4$ [Hz]
32 kbps	–	–	–	–
64 kbps	112	117	142	192
96 kbps	81	114	140	118
128 kbps	71	98	123	106

TABLE VIII
MEAN VALUES OF FORMANT DIFFERENCES FOR OGG CODEC WITH DIFFERENT BITRATE VALUES

OGG	$\Delta F1$ [Hz]	$\Delta F2$ [Hz]	$\Delta F3$ [Hz]	$\Delta F4$ [Hz]
32 kbps	112	117	142	192
64 kbps	39	77	96	113
96 kbps	34	60	69	113
128 kbps	15	27	36	65

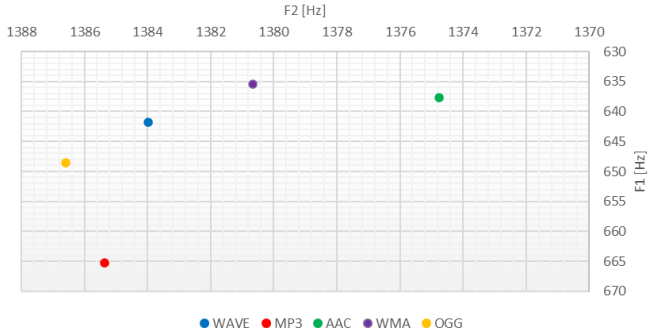

Fig. 8. Example of a formant triangle for first two formants of a vowel ”a” for Japanese speech

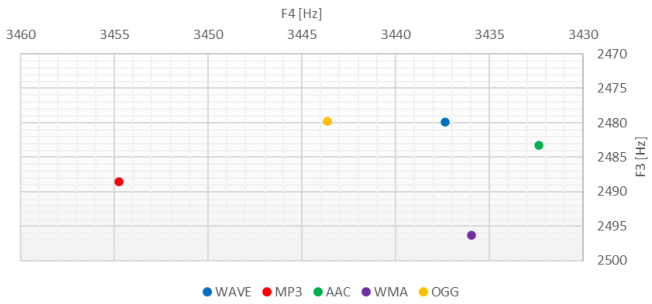

Fig. 9. Example of a formant triangle for second two formants of a vowel ”a” for Japanese speech

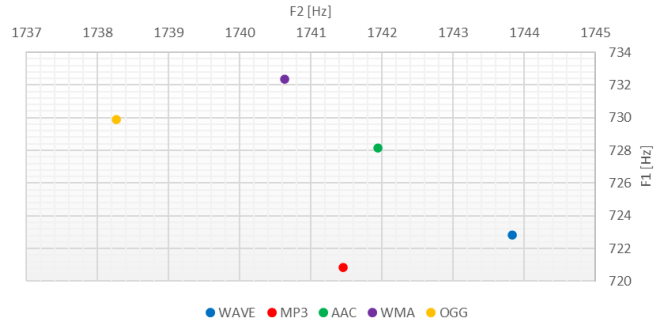

Fig. 10. Example of a formant triangle for first two formants of a vowel ”a” for English speech

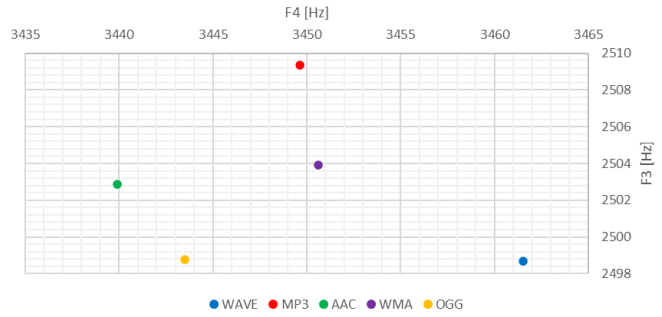

Fig. 11. Example of a formant triangle for second two formants of a vowel ”a” for English speech

III. RESULTS AND DISCUSSION

Impact of all-purpose lossy codecs AAC, MP3, WMA and OGG, as well as speech- focused Speex on third and fourth formant frequencies was examined. Additionally, we checked the impact of whether the speaker is or is not a native Japanese speaker. Similarly native English speakers were checked, but with a different dataset. Also impact of bitrate on differences between originals' and encoded files' format frequencies was checked. All four previously examined codecs mostly follow the same pattern, as the difference between formant frequency value extracted from an original file and an encoded one rise with the number of the formant. For example, for the files with 16 kHz sampling frequency that were encoded using OGG codec value of the difference for the first formant frequency is 36 Hz, for the second 40 Hz, third 49 Hz and fourth 50 Hz. In case of Japanese language the differences seem to be unaffected by the nationality of the speaker, however the database available for the author was limited, and this aspect might need further investigation. For male English speakers received results are similar or with slightly smaller differences, however it must be pointed out that in this case only single vowel was examined. For female English speakers MP3 and OGG codecs manifested similar results to their male counterparts, however AAC and WMA displayed noticeably bigger differences. A speech-focused codec Speex was also

examined in the same manner as all-purpose codecs before. It does seem to follow the same pattern, as well for all first four formant frequencies. The overall difference values are significantly higher than those of any all-purpose codec with the values ranging from 58 Hz up to 114 Hz, however, it must be noted that files encoded with this codec had significantly smaller bitrate values. Finally, the bitrate value seems to have different impact on the difference value, depending on the codec itself. For OGG and WMA codecs the decrease of bitrate value increases difference values significantly. This is also true to some extent for MP3 codec, however, for the AAC codec it doesn't seem to have any impact. The results presented in this paper might provide additional information for a number of applications, including forensic science, where accuracy of speaker recognition and verification are crucial. They may also provide valuable information for speech recognition using neural networks as they tend to utilize spectrogram imagery that will also be influenced by errors introduced by lossy coding.

REFERENCES

[1] H. Tachibana and Y. Suzuki, " "Acoustical Science and Technology" – An improved version of the "Journal of Acoustical Society of Japan (E)" –," *Acoust. Sci. & Tech.*, 22, 1–1 (2001).

[2] L. G. Kersta, (1962), *Voiceprint Identification*, Nature, 196, 1253 - 1257

[3] Y. Kinoshita, (2001) *Testing Realistic Forensic Speaker Identification In Japanese: A Likelihood Ratio Based Approach Using Formants*, PhD Thesis, Australian National University

[4] H. Hollien , R. Schwarz (2000), *Aural-perceptual speaker identification: Problems with noncontemporary samples*, Forensic Linguistics: The International Journal of Speech, Language and the Law, 7, 2, 199-211

[5] S. Brachmański , (2015) *Selected problems of speech transmission quality assessment* (in Polish – *Wybrane zagadnienia oceny jakości transmisji sygnału mowy*), Wroclaw: Oficyna Wyd. Politechniki Wroclawskiej

[6] M. Kucharski, S. Brachmański, (2019) *Coding Effects on Changes in Formant Frequencies in Japanese Speech Signals*, Vibrations in Physical Systems, 1, 30, 243-250

[7] ITU-T Recommendation P.501, (2017) *Test signals for use in telephonometry*

[8] ITU-T Recommendation P.800, (1996) *Method for subjective determination of transmission quality*

[9] M. Kucharski, (2017) *Realization of Japanese sentences sets acoustical database for selected coding techniques*, Wroclaw, BSc Thesis, Wroclaw University of Science and Technology

[10] Y. Hirata, K. Tsukada, (2004) *The Effects of Speaking Rates and Vowel Length on Formant Movements in Japanese* Proceedings of the 2003 Texas Linguistic Society Conference

[11] T. Hirahara, R. Akahane-Yamada, (2004) *Acoustic Characteristics of Japanese Vowels*, 18th International Congress of Acoustics

[12] P. Warden, (2018) *Speech Commands: A Dataset for Limited-Vocabulary Speech Recognition*, http://download.tensorflow.org/data/speech_commands_v0.01.tar.gz access 22.09.2020

[13] J. M. Valin (2007) *The Speex Codec Manual Version 1.2 Beta 3* Xiph.org Foundation

[14] https://docs.microsoft.com/en-us/windows/win32/medfound/about-the-windows-media-codecswindows-media-audio-9 (access 15.07.2020)