Coding Effects on Changes in Formant Frequencies in Japanese Speech Signals

Mateusz KUCHARSKI
Wroclaw University of Science and Technology, Faculty of Electronics, Department of Acoustics and Multimedia, 50-370 Wroclaw, Wybrzeze Wyspianskiego 27, 223513@student.pwr.edu.pl

Stefan BRACHMAŃSKI
Wroclaw University of Science and Technology, Faculty of Electronics, Department of Acoustics and Multimedia, 50-370 Wroclaw, Wybrzeze Wyspianskiego 27, stefan.brachmanski@pwr.edu.pl

Abstract

This paper presents results of research on effects of lossy coding on formant frequencies for japanese speech signals. Additionally changes in pitch of the voice were inspected. For this research four most popular lossy coding standards were chosen, MP3, WMA, AAC and OGG, and compared to original WAVE files. Audio files were created by the author based on ITU-T P.501 recommendation in two sampling frequencies, 16 kHz and 48 kHz, and converted into chosen codecs. To extract the data from audio files, open license software Praat was used. Due to discovered differences in time duration between original and encoded files, that also differed between individual codecs, only OGG and WMA standards were compared directly. MP3 and AAC standards were divided into Japanese syllables, averaged and then compared into also averaged WAVE files. Results were additionally compared to FLAC lossless codec.

Keywords: speech, speech coding, formant

1. Introduction

Researching coding influence on formant frequencies requires creating database of speech signals. Database of speech signal used for this research was created by the author as a part of BSc thesis (Tab. 1) [1, 4]. Sentences used were taken from ITU-T P.501 recommendations sentence list of Japanese language [2]. They were recorded in environment meeting the requirements of ITU-T P.800 using condenser microphone and open licence software [3]. Sampling rate during recording was 48 kHz, later additional 16 kHz audio files were created using those original files. All files were monophonic and coded in 16 bit PCM. Recordings of the same sentences were made twice, during two different sessions. Original WAVE files, in both 16 kHz and 48 kHz, were later converted into four chosen most popular lossy codec formats: MPEG Layer 3 (MP3), Windows Media Audio (WMA), Advanced Audio Coding (AAC) and OGG Vorbis (OGG), as well as lossless FLAC. Final database consists of 192 audio files.

During preparations for the experiment, several issues were discovered. Before data extraction begun, files' parameters were checked by additional software and during that author found out differences in time durations between original WAVE and encoded files. Those encoded with MP3, AAC and WMA codecs had longer time duration than

originals, while those with OGG and lossless FLAC were identical in that matter. The additional lengths were also different depending on the codec. Differences were smallest for the WMA codec, 5 ms for the first file, and noticeably larger for MP3, reaching 85 ms. The AAC codec was also problematic, even during database creation, and due to inconsistent bitrate used software was unable to identify its length. While partly responsible for the differences turned out to be codec starting in the beginning of the file, it was only about 15 ms, and the remaining 70 ms were in the file itself. The origin of these differences is unknown.

Table 1. Japanese sentences contained in database.
The transcription of Japanese texts come from ITU-T.501 recommendation

Number of sentence	**Notation**	**Sentence**
1	Original	彼は鮎を釣る名人です。
	Hiragana	かれ わ あゆ を つる めいじん です。
	Transcryption	Kare wa ayu wo tsuru meijin desu.
2	Original	古代エジプトで十進法の原理が作られました。
	Hiragana	こだい えじぷと で じゅっしんほう の げんり が つくられました。
	Transcryption	Kodai ejiputo de jusshinhou no genri ga tsukuraremashita.
3	Original	読書の楽しさを知ってください。
	Hiragana	どくしょ の たのしさ を しって ください。
	Transcryption	Dokusho no tanoshisa wo shitte kudasai.
4	Original	人間の価値は知識をどう活用するかで決まります。
	Hiragana	にんげん の かち わ ちしき を どう かつよう するか で きまります。
	Transcryption	Ningen no kachi wa chishiki wo dou katsuyou suruka de kimarimasu.
5	Original	彼女を説得しようとしても無駄です。
	Hiragana	かのじょ を せっとく しようとしても むだです。
	Transcryption	Kanojo wo settoku shiyoutoshitemo mudadesu.
6	Original	その昔ガラスは大変めずらしいものでした。
	Hiragana	その むかし がらす わ たいへん めずらしい ものでした。
	Transcryption	Sono mukasi garasu wa taihen mezurashii monodeshita.
7	Original	近頃の子供たちはひ弱です。
	Hiragana	ちかごろ の こども たち わ ひよわ です。
	Transcryption	Chikagoro no kodomo tachi wa hiyowa desu.
8	Original	イギリス人は雨の中を平気で濡れて歩きます。
	Hiragana	いぎりすじん わ あめの なかを へいきで ぬれて あるきます。
	Transcryption	Igirisujin wa ameno nakawo heikide nurete arukimasu.

Another problem that occurred was that software used for data extraction, program Praat, which did not support OGG, AAC and WMA files. To bypass this issue, files in these encodings were converted once again into WAVE. Because they were encoded with lossy codecs, the changes in parameters, that author was interested in, should have already impacted those files and thus be identical after another conversion using a lossless encoding. Also because AAC format is primarily designed to contain stereo data, it artificially created double tracks from original mono. Thus it was necessary to convert it back to mono using Audacity software.

2. The research

The problem of lengths was resolved by using two methods of comparing files. First method was used for OGG, WMA and FLAC codecs, as OGG and FLAC had the same number of points in which the formant measurements were taken and the WMA codec only differed by several points. The method was a simple side-by-side comparison. Using Praat software, a list of formant points was extracted, with a 0.00625 second interval between each point, for both the examined codec and original WAVE file. Than a simple subtraction for the first and second formant frequencies in each point between two files was conducted, and an absolute value of the result was taken.

The second method was a bit less straightforward. As the differences in formant points were too big, files could not be compared side-by-side. It was decided that each file was going to be divided into several parts, then data from each part would be extracted separately. A mean value would be taken and compared to mean value of corresponding part of the WAVE file. Fact, that Japanese language was being examined during this research turned out to be important in this part. Words might have been too long as chunks of data, and also a word can be really short as well as really long. On the other hand singular letters can be extremely hard to separate from each other. Fortunately, Japanese syllables are quite easy to separate. Also unlike western syllables, there is a very well defined number of them, due to writing systems functioning in Japan. There are three writing systems there:

- Kanji - Chinese logographic characters, that are used for the words' meanings (there are about 2200 of basic characters and many more specialistic and archaic ones);
- Hiragana - syllabary consisting 46 basic characters, 25 diacritics and 36 digraphs, used mostly for particles and word ending modifications;
- Katakana - syllabary consisting of the same syllables that hiragana, but used for foreign-borrowed words.

All of Japanese language can be written in both syllabaries, and it often is, for children and people learning it, so it was decided that Japanese syllables were going to be those needed chunks of sentence for the research.

The data was extracted using "view and edit" mode in Praat software. This is the mode that displays time course on a graph in the upper and spectrogram in the lower part of the screen, both of which share the same time axis. It is also possible to display formant points (red dots) as well as pitch (blue line) on the spectrogram. This display, with aid of playing selected parts of examined file, enabled the author to fairly

accurately separate syllables and thus extract wanted data. So, from each chunk, the list of formant points was extracted, and then mean values for first two formants were calculated. Additionally maximum and minimum values were noted.

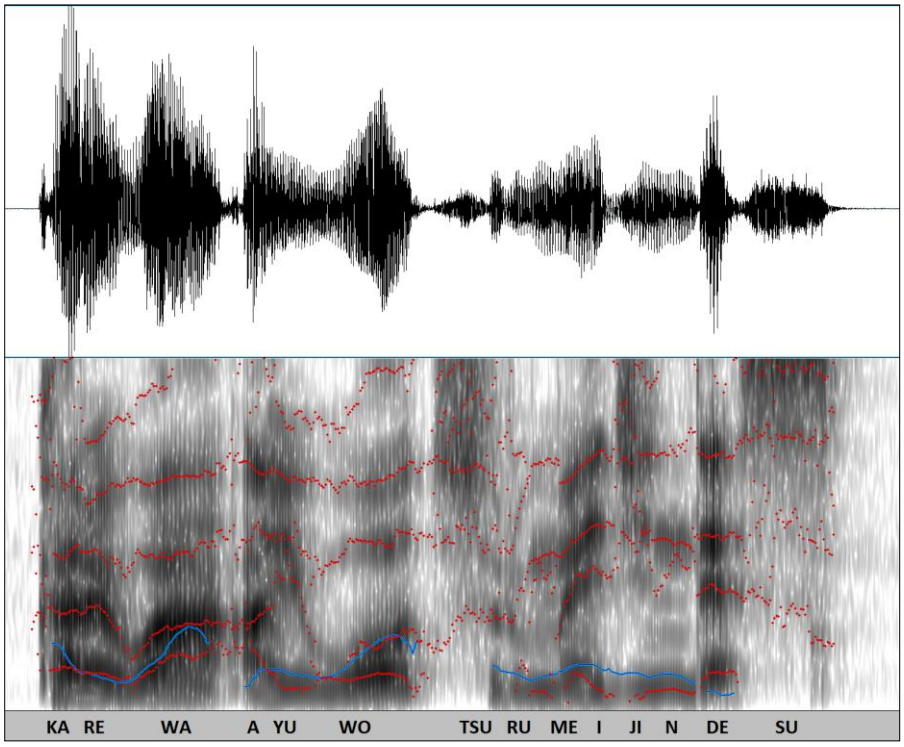

Figure 1. Time course (upper part) and spectrogram with formants and pitch (lower part) for the first sentence

In similar fashion data about pitch of the voice was collected. For OGG, WMA and FLAC codecs it was extracted as a list for the whole file. For the rest, it was extracted from the same parts (syllables), as formant data. There, the mean value as well as minimum and maximum were taken from software.

Data collected for each codec was then compared to corresponding data from original WAVE files. For the presentation, absolute values of differences were calculated. Results of this are presented in tables 2 and 3. There is one table containing mean values of differences in formant frequencies and their standard deviations for all lossy codec, as well as a similar table containing data about pitch, also for all codecs. Each table contains results for two versions of the sentence, recorded during two different sessions. Each version is also available in both 16 kHz and 48 kHz sampling frequency.

The comparison of lossless FLAC to original WAVE files resulted in exactly none differences - all formant points' values were the same for both. This result was expected, but its' point was not to check if there are any differences, but if the software used for

this research does not influence the outcome. Because differences were nonexistent, with respect to this standpoint, it is safe to assume that results achieved for the rest of researched files are accurate. As the results for this part were all zeros, it was assumed that presenting them in a table is not necessary.

Table 2. Mean values of formant differences with standard deviation for all codecs

Codec	Recording 1				Recording 2			
	16 kHz		48 kHz		16 kHz		48 kHz	
	F1	F2	F1	F2	F1	F2	F1	F2
OGG	35.51	40.27	15.68	17.43	34.18	35.12	13.85	16.51
Standard deviation	12.27	6.62	7.31	14.33	17.70	8.86	10.30	14.46
WMA	50.39	59.66	65.35	81.71	60.35	67.32	60.88	76.89
Standard deviation	12.06	13.28	17.43	23.18	29.75	24.88	11.59	18.85
MP3	53.27	61.19	48.43	53.59	34.78	52.83	29.49	35.70
Standard deviation	17.81	32.25	11.05	16.76	8.62	42.24	7.26	6.02
AAC	40.41	52.03	45.79	64.76	38.59	59.61	33.59	43.44
Standard deviation	18.26	31.56	20.11	41.01	11.00	50.19	8.52	8.71

Table 3. Mean values of pitch with standard deviation for all codecs

Codec	Recording 1		Recording 2	
	16 kHz	48 kHz	16 kHz	48 kHz
OGG	0.40	0.10	0.34	0.01
Standard deviation	0.60	0.25	0.35	0.01
WMA	2.59	3.83	5.09	5.88
Standard deviation	1.79	1.88	4.78	4.75
MP3	2.73	1.82	6.48	3.80
Standard deviation	3.74	0.86	9.10	4.86
AAC	2.92	4.51	2.31	3.36
Standard deviation	3.59	5.10	2.62	3.46

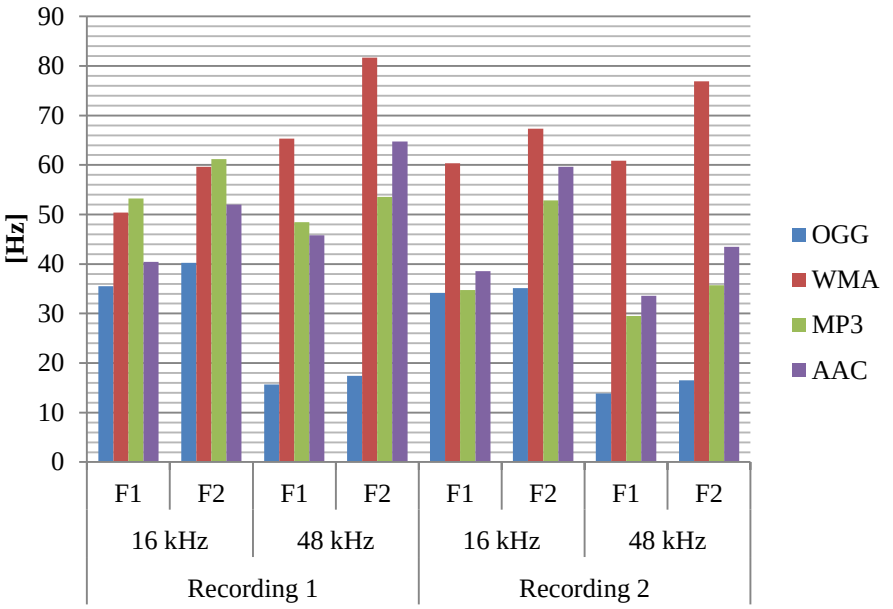

Figure 2. Mean values of formant differences for all codecs

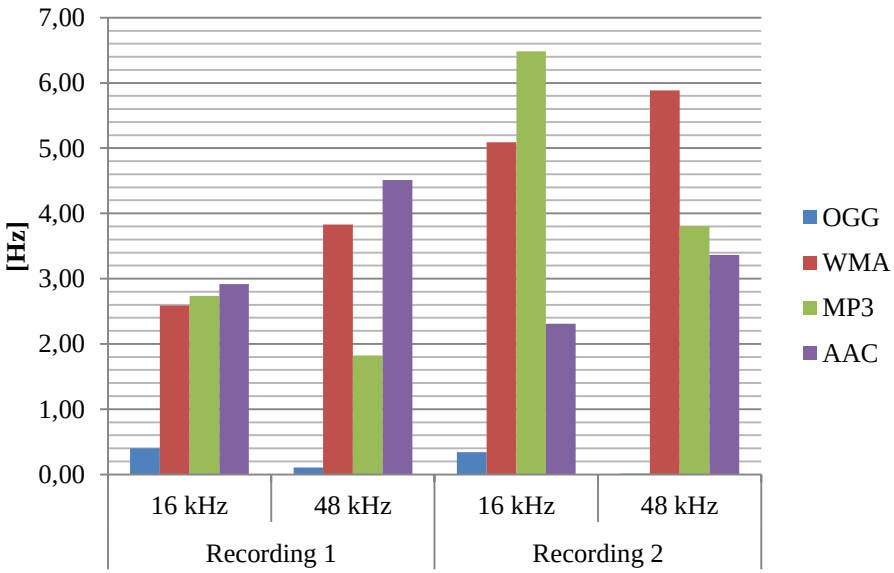

Figure 3. Mean values of pitch for all codecs

3. Results

Differences in formant frequencies fluctuate from about 15 Hz up to 80 Hz, depending on codec, sampling frequency and number of the formant. The lowest values were achieved for the OGG codec. Recording session did not matter that much in this case, however there is a significant difference between sampling frequencies. While for the original 48 kHz results are quite small, about 15 Hz, the converted 16 kHz versions have over twice as much. In WMA codecs' case the values are much higher, in fact they are the highest among all codecs. It might be caused by the method used for data extraction. It was examined the same way OGG was, but unlike OGG there are differences in lengths. They are not very big - highest noted difference in number of formant points was six and the mean value is about three. Also sentence number five has unusually high values, around and above 100 Hz difference, with 120 Hz being maximum for the first formant and 124 Hz for the second. MP3 codec was the first to be examined using the second method, that is by dividing file into parts. Mean values are lower than WMAs', but it also has some spikes over 100 Hz, reaching even 183 Hz in one point. Those points however are exceptions and the rest are close to mean value. The AAC codecs' results are very similar to MP3s', they even have spikes in the same places and with similar values.

Results of research in terms of pitch of the voice are also interesting, because there is almost no difference between original file and lossy encoded ones. As it can be seen in tables above, OGG has again the lowest differences, which are almost always below 1 Hz. The rest of the codecs give a few hertz differences with mean values usually not exceeding 6 Hz. Again, there are several files that have higher difference values, up to 14 Hz. In case of WMAs' higher values, it is also fault of uneven lengths. It can be seen that generally the values are also below 1 Hz, but there are additional points, just like formant points, that add values above 100 Hz. Because of that mean values are higher. There were also several unusual occurrences, when software detected pitch on much higher than normal frequencies, from 400 up to 500 Hz. These are however rare and do not last long, so it is safe to assume that in general lossy coding has almost no effect on pitch of the voice.

4. Conclusion

The objective of conducted research was to check, if popular lossy codecs have effect on certain voice parameters. The parameters were formant frequencies and pitch of the voice. Results presented in this paper show that for the formant frequencies there are in fact some differences. OGG codec, that was the easiest to examine, had differences of about 15 Hz for 48 kHz sampling frequency, and about 35 Hz for 16 kHz. The rest of the codecs, WMA, MP3 and AAC, had much higher values, with WMA having the highest, above 60 Hz mean, up to even 120 Hz. MP3 and AAC had slightly lower overall values, but in case of one file got almost up to 190 Hz. Unlike formant frequencies, pitch of the voice had almost no changes at all, again with several exceptions.

There are several potential perspectives for the future expanding of this research. It is possible to expand the database for other speakers, including natives. It is also possible

to check the results for another languages and research influence of the language itself on those parameters. Also another, more accurate methods might be developed.

References

1. S. Brachmański, *Wybrane zagadnienia oceny jakości transmisji sygnału mowy*, Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej, 2015.
2. ITU-T Recommendation P.501, *Test signals for use in telephonometry*, 2017.
3. ITU-T Recommendation P.800, *Method for subjective determination of transmission quality*, 1996.
4. M. Kucharski, *Realization of Japanese sentences sets acoustical database for selected coding techniques*, Wrocław, BSc Thesis, Wrocław University of Science and Technology, 2017.