A PRELIMINARY STUDY ON THE INF0LUENCE OF SOUND DATA COMPRESSION UPON FORMANT FREQUENCY DISTRIBUTIONS IN VOWELS AND THEIR MEASUREMENT

Bogdan Rozborski

Polish-Japanese Institute of Information Technology
b.rozborski@chello.pl

ABSTRACT

The aim of this paper is to demonstrate the appearance of spectral differences of formant structures of a chosen vowel that occurred after compressing PCM sound data using a given compression method. Experiments show that sound data compression does not significantly affect stability of formant frequency distributions in terms of statistics, as long as it does not introduce a random, stochastic component into the original speech signal.

Since used in experiments the formant frequency measuring technique, based on the Burg's LPC algorithm, is vulnerable to a noise component in the speech signal, it becomes crucial to estimate the contribution of both, systematic and random errors, introduced by signal compression techniques, to the overall formant frequency measurement error. Bearing in mind that the sound data compression gains more popularity nowadays, and that formant frequency analysis is one of the basic tools for speaker identification in forensic practices, it is essential to know what the chances are for correct speaker identification while dealing with criminal evidence in the form of a compressed sound file.

Keywords: Sound data compression, formant frequency analysis, speaker identification.

1. INTRODUCTION

The speaker identification is a criminological task that imposes a high degree of responsibility on the part of an expert, hence it cannot be performed with a use of a single test. Speaker identification utilizes tests based on three fundamental domains characterizing human speech production and perception, namely: acoustics, physiology and linguistics. One of the tests performed by a forensic expert that involves all three domains is the analysis of formant frequency distributions. Indeed, as noted in [2], a formant represents a local spectral maximum occurring at a certain resonance frequency. Configurations of several formants determine the quality (both phonological and personal) of speech sounds, and relations among traces of formants along time axis show articulatory relations between consecutive speech sounds.

Due to technical limitations that apply to spontaneous speech recordings, the speaker identification procedure, based on the formant analysis, usually involves the lowest three formants that characterize vowels, the most prominent (containing the highest amount of acoustic energy) speech sounds. In order to preserve and emphasize personal differences/similarities characterizing compared speech samples, an expert builds vowel sets consisting of vowels of the same spectral quality, taken from a precisely defined consonantal context. These sets of vowels may be thought of as spaces composed of several dimensions that represent formants or in statistical terms as random samples with several variables.

Being aware of the fact that compared speech samples may be recorded using different techniques (including sound data compression), an expert should be interested in revealing spectral differences of formant structures characterizing compared vowel sets in order to apply an appropriate correction for formant frequency measurements.

2. METHODOLOGY

The analyzed data sets were drawn from vowel sets obtained from speech samples representing various compression techniques, and from the reference vowel set that comes form the original PCM sound. Due to the complexity of the discussed problem, only comparisons between the reference data set and the compression data sets were done. The presence or lack of the difference between compared data sets were expressed qualitatively by applying T-tests and calculating cross-correlation factors for each variable (formant). However, an attempt was made to present a ranking of compression methods depending on the amount of change they introduce to the original speech signal.

2.1. Original speech signal

As a reference speech signal the author used a professional studio recording of controlled speech produced by a male actor who read a passage of a Polish text. The 90 seconds long piece of recording was extracted form a standard CD recorded in the uncompressed PCM format (cf. [3]), at sampling frequency of 44,1kHz and resolution of 16 bits.

2.2. Test speech signals

Tested speech signals were obtained form the reference speech signal recording by converting the latter into ten different sound files using the following seven most popular compression methods:

- ATRAC (Adaptive Transform Acoustic Coding) is an audio coding system based on psychoacoustic principles. (cf. [7, 4])
- Dialogic ADPCM (Adaptive Differential Pulse Code Modulation) is an audio compression scheme which compresses 16-bit audio data into 4-bit audio data
- GSM 06.10 is a compression method used in mobile telecommunication, based on Regular Pulse Excited - Linear Predictive Coder (RPE-LPC) with a Long Term Predictor loop. Speech signal is transmitted with a total bit rate of 13 kbit/s (cf. [8])
- GSM "real" refers to a sound file created with the use of hardware (cellular phones). As confirmed by the author's phone operator, the bit rate of speech data transmission is also 13 kbit/s
- MP3 (MPEG Layer-3) is another audio coding system based on psychoacoustic principles (cf. [1])
- U-LAW is an encoding format used in telecommunication that compresses original 16-bit audio down to 8 bits with a dynamic range of about 13 bits (cf. [9])
- WMA (Windows Media Audio), proposed by Microsoft, is an audio coding system based on psychoacoustic principles (cf. [1])

The table below lists the set of compressed sound files used in further analysis. Four files created in MP3 format were produced to demonstrate the dependence between compression rate and the amount of spectral change introduced to the original PCM data by the sound compression.

Table 1: The list of compressed speech signal files used in the analysis.

sound file mnemonic	compression applied	compression rate[1] cr	Compression factor cf [%]
ATRAC	ATRAC	5:1	3,48
ADPCM	Dialogic ADPCM	4:1	2,32
GSM	GSM 06.10	54,3:1	60,67
rGSM	GSM "real"	54,3:1	60,67
MP3 12,6	MP3	12,6:1	12,30
MP3 35,3	MP3	35,3:1	38,63
MP3 44,1	MP3	44,1:1	48,84
MP3 88,2	MP3	88,2:1	100,00
U-LAW	U-LAW	2:1	0,00
WMA	WMA	20,7:1	21,69

Shown in the table 1 compression factor denotes a relative compression rate. This factor will be used to determine a compression efficiency coefficient (cec) applied later for ranking of tested compression methods. All factors used for estimation of the cec coefficient are expressed in percent points. The compression factor is defined as follows:

$$cf = \frac{(cr - MIN(cr))}{(MAX(cr) - MIN(cr))} * 100 \qquad (1)$$

The next step was decoding compressed files back to the uncompressed PCM sound files used for the formant frequency measurements.

2.3. Data preparation

A low, back vowel [a] was extracted from each speech signal sample to produce analytical material. The most reliable formant frequency measurements are obtained for vowels with maximal distances between formant frequencies, and high signal to noise ratio. The [a] vowel, as opposed to high or round, back vowels [i] or [u] guarantees relatively distant formants' placement on the frequency axis, and it is the most sonorous speech sound in Polish, giving a high signal to

[1] Compression rate cr is defined as a ratio of uncompressed PCM audio data bit rate to the bit rate of compressed audio data. The bit rate of uncompressed PCM audio used by the author is 705,6 kbit/s.

noise ratio. Another reason for choosing [a] vowel was relatively high frequency of its occurrence in Polish language. [a] vowels were selected from a consonantal contexts excluding palatal and nasal consonants. Those [a] vowels that occurred before or after other vowels were discarded too.

The vowel extraction procedure produced eleven vowel sets representing each speech sample. Each vowel set contains thirty seven instances of [a] vowel. All thirty seven instances are exactly the same across all eleven vowel sets (i.e. vowels were taken from the same consonantal context in all vowel sets).

A preliminary spectral analysis of all vowel sets revealed that only the three first formants could be measured with relatively small error. Higher formants were cut off or much distorted by the noise component. Formant frequency measurement procedure was conducted automatically, with time step of 0,00625s, over a whole vowel set, with the duration of 2,1s, giving 336 formant measurements. This means that the formant frequencies were measured over whole segments, and the number of measurements depended on the segment length. The Burg's algorithm implemented in the Praat program (cf. [6]) was used as a measuring method. As an outcome of this procedure, eleven data sets were produced. Each data set consists of 336 formant measurements. Each measurement contains three values representing three formant frequencies.

2.4. Data analysis

Since a vowel articulation is not a stable process in time (organs of speech move while pronouncing a vowel), and the speech production is affected by various "disturbing" psycho-physiological phenomena like neural noise (cf. [4, 5]), the formant frequencies are not equal to one particular value, but they form a distribution around a particular mean frequency. Because of the random, stochastic nature of those speech disturbing psycho-physiological factors, empirical formant frequency distributions may be approximated with normal/Gaussian distribution. It follows from the above that the statistical methods shall be used to account for the differences between the reference speech sample and samples of compressed signals.

The following statistical procedures were used for the analysis. As a normality test conducted on the empirical formant frequency distributions, the Shapiro-Wilk normality test was applied to all three formants in all eleven data sets. All thirty three tests gave a positive result (W values were greater than critical Wα = 0,989 for n = 300 and significance level α = 0,05). This means that the data collected for the experiment are statistically valid and statistics based on a normal/Gaussian distribution may be applied to the data sets.

In order to determine whether the differences between the mean values of compared formant frequency distributions are statistically significant, the T-tests were applied with a probability threshold of 0,05. The null hypothesis is that compared mean frequencies are equal and characterize the same population (compared speech samples come from the same source/speaker), and their empirical differences are not statistically significant. It should be also noted that in the preliminary analysis, the data sets were not tested for equal/not equal variances.

To estimate the degree of dependence between compared formant frequency distributions (uncompressed vs. compressed) the cross-correlation coefficients were calculated for three formants, in all compressed data sets (this gives thirty tests).

3. THE RESULTS

Tables 2 and 3 below demonstrate the results of the experiment. For the sake of brevity however mean values, standard deviations and T-values were not included in table 2.

The following symbols are used in the table: cc - cross correlation coefficient multiplied by 100, tf - "T-test factor" used for the ranking cec coefficient calculation, corf - "cross-correlation factor" defined as an average of cc for a given compression signal sample. The tf factor is defined as an average of T-test probabilities for a number of formants in a compressed signal sample. The tf value is given by the following formula:

(2) $$tf = \frac{1}{fn} * \sum_{1}^{fn} p_i * 100$$

where p_i - T-test probability, fn - a number of formants in a given sound sample. The ranking coefficient cec is defined as:

(3) $$cec = \frac{(cf + corf + tf)}{3}$$

A closer look at the table 2 below makes it possible to conclude that all compression methods introduce certain changes to the formant frequency distributions. These changes affect both mean values and variances, which suggests that compressing speech signal using a given method introduces both a systematic and random change factor into the frequency spectrum of an original speech signal. It is also evident that exceeding the compression rate (cf. the case labeled as MP3 88,2, in table 2 below) may be as destructive to original formant frequency distributions as introducing noise while speech signal transmission via the GSM telephony system (cf. the case rGSM from the table 2 below).

Table 2: The results of statistical comparison of formant frequency distributions for three formants in ten examples of compressed speech signal.

	F1		F2		F3		factors	
mnemonic	pi	cc	pi	cc	pi	cc	tf	corf
ATRAC	0,57	96,07	0,75	98,94	0,69	99,61	67,23	98,21
ADPCM	0,00	89,49	0,89	97,68	0,48	96,90	45,69	94,69
GSM	0,71	91,73	0,71	96,24	0,00	91,64	47,56	93,20
rGSM	0,00	53,49	0,20	97,56	0,00	84,30	6,73	78,45
MP3 12,6	0,96	98,73	0,83	98,66	0,10	97,23	62,82	98,21
MP3 35,3	0,87	98,87	0,45	97,38	0,59	96,98	63,56	97,74
MP3 44,1	0,41	98,05	0,53	97,30	0,39	97,62	44,23	97,66
MP3 88,2	0,02	91,31	0,00	94,69	0,00	57,50	0,53	81,17
U-LAW	0,95	99,63	0,86	99,41	0,90	98,72	90,10	99,25
WMA	0,85	98,28	0,65	99,20	0,70	98,50	73,45	98,66

To summarize the discussion it seems worth presenting a tableau with compression methods ranked according to the formula (3).

Table 3: The ranking tableau for compression methods used with speech signal.

sound file mnemonic	Ranking
GSM	67,14
MP3 35,3	66,64
WMA	64,60
MP3 44,1	63,58
U-LAW	63,12
MP3 88,2	60,57
MP3 12,6	57,78
ATRAC	56,31
rGSM	48,62
ADPCM	47,57

Table 3 above shows that the most efficient in terms of compression rate, and amount of changes introduced to the original signal, compression scheme is the GSM 06.10 format designed especially for speech signals encoding. In more general terms, compression techniques that utilize Regular Pulse Excited - Linear Predictive Coder seem to handle the problem of speech signal compression in a quite satisfactory manner. It is also interesting to observe how pure computer simulation of the GSM format is different from application of the very same format in a real cellular phone system.

4. CONCLUSIONS

The analysis of the data presented in this paper show clearly that the amount of change introduced to the original speech signal, by application of a particular compression technique, is by all means statistically significant. There is only a limited number of sound data compression methods, (e.g. the MPEG Layer-3 format with compression rates of 35,5:1 or 44,1:1) whose application to speech signals does not cause statistically significant changes in the original signal. Bearing this in mind, a forensic expert must produce comparative speech samples applying exactly the same compression technique as has been applied for producing an evidence recording. A speech signal analyst should keep in mind that compressing speech signals may introduce statistically significant changes to its spectrum, which contributes to an increase of overall measuring error. Further research is needed however to establish the actual level of systematic and random error in compressed speech signal measurements.

5. REFERENCES

[1] Hacker, S. 2000. *MP3: The Definitive.* Warszawa: Helion.pl.

[2] Jassem, W. *1983. The Phonology of Modern English.* Warszawa: Państwowe Wydawnictwo Naukowe, 140-165.

[3] Lyons, R. 2003. Understanding Digital Signal Processing. Warszawa: Wydawnictwa Komunikacji I Łączności, 37-55.

[4] Maruszewski, M. 1970. *Mowa a mózg.* Warszawa: Państwowe Wydawnictwo Naukowe.

[5] Ozimek, E. 2002. *Sound and Perception, Physical and Psychoacoustic Aspects.* Warszawa-Poznań: Wydawnictwo Naukowe PWN.

[6] Praat: doing phonetics by computer. http://www.fon.hum.uva.nl/praat/ visited 1-March-2007.

[7] Veldhuis, R., Breeuwer, M., van der Wall, R. 1989. Subband coding of digital audio signals without loss of quality. *Proc. International Conference on Acoustics, Speech and Signal Processing,* Glasgow, 2009-2012.

[8] Scourias, J. Overview of the Global System for Mobile Communications. https://styx.uwaterloo.ca/~jscouria/GSM/gsmreport.html#1 visited 4-June-2007.

[9] µ-law algorithm. http://en.wikipedia.org/wiki/%CE%9C-law_algorithm visited 6-June-2007.