The effect of MPEG audio compression on multidimensional set of voice parameters

Julio Gonzalez[1] and Teresa Cervera[2]

From the [1]University Jaume I of Castellon, Castellon, Spain, and [2]University of Valencia, Valencia, Spain

Log Phon Vocol 2001; 26: 124–138

The MPEG-1 Layer 3 compression schema of audio signal, or commonly known as *mp3*, has caused a great impact in recent years as it has reached high compression rates while also conserving a high sound quality. Previous listening tests have shown that music and speech samples compressed at high bitrates are virtually indistinguishable from the original samples, but very little is known about how compression acoustically affects the voice signal. In Experiment 1 the spectral composition of original and compressed speech signals were analyzed by means of the Long-Term Average Spectrum using the Computerized Speech Laboratory (Kay Elemetrics Corp. (Pine Brook, NJ, USA)). In Experiment 2 a set of 29 voice parameters extracted by using the Multidimensional Voice Program of Kay are compared between original and compressed voices at different bitrates. Results show a high fidelity at high-bitrate compressions (128 and 160 kbit per second (kbps)) both in voice parameters and the amplitude-frequency spectrum. Compressions at 64 kbps or lower bitrates introduces substantial modifications in the voice signal, seriously altering parameters related with tremor, amplitude perturbation, noise, subharmonics and voice irregularities, likewise the signal spectrum is altered in its high frequency region.

Key words: voice parameters, MDVP, MPEG, mp3, Long-Term Average Spectrum, voice analysis, speech compression.

Julio Gonzalez, Department of Basic and Clinical Psychology and Psychobiology, University Jaume I of Castellon, 12080 Castellon, Spain. Tel: +34 964 728000. Fax: +34 964 729350. E-mail: gonzalez@psb.uji.es

INTRODUCTION

As new technologies are developed, it is important to gauge their usefulness in voice clinic and research. The work in signal compression has progressed impressively, particularly with regard to the preservation of sound quality in data reduction. In the near future, powerful compression techniques will be useful in the laboratory to store or transmit voice signals to other laboratories provided that there is no alteration of voice signal properties relevant to clinical and research practice.

In recent years a revolutionary signal compression technique has caused a great impact in the field of the sound and music. The development of the Moving Pictures Expert Group (MPEG) standards in audio coding has achieved very high rates of compression while preserving the high quality of the sound, particularly the most powerful format, MPEG-1 Layer 3, commonly known as mp3—see Brandenburg and Stoll (3) and Brandenburg (2).

The MPEG is a group of experts that work under ISO—the International Standards Organization—to generate standards for digital video and audio compression. One of their main goals is the development of compression algorithms that preserve as much sound quality as possible even at very low bitrates, or total number of bits per second that will be contained in the encoded file. In general, the higher the bitrate, the higher the quality of the sound, but the larger the file will be. Currently, the work of this group, which is carried out by several organizations, is defined by four phases: MPEG-1 or "Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 MBits/s", MPEG-2 or "Generic Coding of Moving Pictures and Associated Audio", MPEG-3, which has been merged into MPEG-2, and MPEG-4 or "Very Low Bitrate Audio-Visual Coding", now in progress. The MPEG-1 includes a family of three audio coding schemes (Layer 1, 2, 3) with increasing encoder complexity and performance. In audio compression, MPEG-2 is only useful for appli-

 ISSN 1401-5439

cations with limited bandwidth, 11250 Hz at best. For applications with full bandwidth, MPEG-1 Layer 3—or mp3—reaches the best sound quality of all codecs.

MPEG-1 Layer 3 is an international ISO/MPEG standard—ISO/IEC 11172-3 (10)—based on a psychoacoustic model that for medium and high bitrates, such as 120 kbit per second (kbps) or more per channel, achieves a very high quality sound. At these bitrates trained listeners found it difficult to detect differences between original and compressed signal. At lower bitrates, Layer 3 is the only audio coding schema that has been recommended by the International Telecommunications Union (ITU-R) for use at 60 kbps per channel.

MPEG-1 Layer 3 is a sub-band coder that applies psychoacoustic coding schemes, removing parts of the signal that are perceptually irrelevant. It divides the signal frequency spectrum into 32 sub-bands matching the psychoacoustic properties of the human ear. For each sub-band an algorithm calculates the perceptual masking effect caused by the other sub-bands. The masking effect raises the threshold of the noise floor, reducing the effective dynamic range of the signal. This reduced range requires less bits for codification and this is the main cause of signal compression. For example, if in the sub-band n the acoustic dynamic range is 60 dB (codified by 10 bits), but the coder calculates the masking effect and finds that any sound 40 dB below is not actually heard, then the effective dynamic range of that sub-band is lowed to $60 - 40 = 20$ dB, codified just by 4 bits. Moreover, the masking effect is computed not only when it is concurrent, but the *mp3* coder also estimates the masking effect that occurs before (2–5 ms) and after (up to 100 ms) a strong sound. This data reduction allows a major compression to store or transmit audio signals without loss of sound quality.

To attain compact disc (CD) audio quality, the audio signal needs to be sampled 44100 times per second and each sample requiring a resolution of 16 bits; this gives 705 kbps, or 1410 kbps if stereo. Listening tests (11, 6) show that practically the same sound quality is obtained with MPEG-1 Layer 3 at 96 kbps and gives a compression ratio of 8.3:1 per channel. For more demanding musical pieces such as piano concerts etc., it is advisable to increase the bitrate to 120 kbps which gives a compression ratio of 6.3:1. Such high compression rates that nonetheless maintain high sound quality have made an enormous impact on the storage and transmission of music via Internet.

The perceptive efficiency discussed here is not exclusive to music; given the results of the listening tests, it can be applied to the speech signal. It could be argued that at bitrates of 96 kbps or more and in many cases even at 64 kbps, the compressed voice is audibly indistinguishable from the original. Apart from some perceptive aspects, we still lack precise information on the degree of distortion that such compression techniques have on the signal. Any form of compression will have some degree of signal modification—though it may not be perceived—that can be reflected by modifications of objective parameters calculated from the voice signal.

The first aim of the study presented here is to ascertain, by way of the Long Term Average Spectrum (LTAS), the extent to which the spectral composition of the voice signal is affected by MPEG-1 Layer 3 compression at different bitrates. The second and major aim is an analysis of how compression at different bitrates affects a set of 29 acoustical voice parameters.

EXPERIMENT 1

Method

Speakers. Subjects were 7 native speakers of *Valenciano* (Catalan) of both sexes (4 females and 3 males), students at the University Jaume I of Castellon (Spain), with ages ranging from 22 to 31. Each exhibited normal speech and audition.

Apparatus. The recording of speech samples was performed with a Shure SM58 microphone at a distance of about 15 cm from the mouth, and a Sony-TCD D-8 digital audiotape (DAT) recorder with a sample frequency of 44.1 kHz. The DAT recorder also can record at 48 kHz but a sample frequency of 44.1 kHz was chosen as this is the optimal output frequency for all the compression conditions applied, with the exception of the 32 kbps compression. The voice samples were analyzed on the Computerized Speech Lab (CSL) Model 4300 developed by Kay Elemetrics Corp. (Pine Brook, NJ, USA).

Voice samples. Each subject read a total of 18 sentences written in *Valenciano* (Catalan) in his natural voice and at normal speed. The six central sentences were used to carry out the acoustic analysis. All voice samples were recorded in a soundproof room at the University Laboratories.

Each voice sample was directly introduced from the DAT recorder to a CSL model 4300 on a PC Pentium at 166 MHz. Then all voice signals were compressed by means the Fraunhofer-Thomson compression scheme at the following bitrates and sample frequencies: 160 kbps (44.1 kHz), 128 kbps (44.1 kHz), 96 kbps (44.1 kHz), 64 kbps (44.1 kHz), 48 kbps (44.1 kHz), and 32 kbps (22050 Hz). These

values give a set of compression rates ranging from 4.4:1 to 22.1:1.

Acoustic analysis. Each original and compressed voice sample was analyzed with the CSL model 4300 performing an LTAS on the amplitude values measured in decibels. The following values were selected for the analysis: a frequency range display of 0–12 kHz, no pre-emphasis, frame size: 5.8 ms (256 points), window weighting: Hamming. The analysis brought up amplitude values at 172.2 Hz intervals, obtaining a total of 71 measurements for each speaker and corresponding to the following frequencies: 0, 172.2, 344.53 . . . and so on up to 12058.59 Hz. The compressed signal at 32 kbps, which had a sample frequency of 22050 Hz, was analyzed at a frame size of 128 points in order to maintain the same interval between consecutive values. Its frequency range encompassed 65 different values: 0, 172.27, 344.53 . . . right up to 11025 Hz.

Results

For each compression (and original) condition, the spectrum average across subjects was calculated. Fig. 1 shows the mean amplitudes in the 0–12 kHz range. To avoid negative values, all data were shifted up by 20 dB. We can observe that up to the point of 6800 Hz, the seven lines that are plotted are close together and parallel. At this frequency, the sample compressed at 32 kbps starts to diverge away from the others, as there is a drastic fall in its energy level. This marked decline is basically due to compression; even though we are dealing with a signal with a lower sample frequency (22050 Hz), its normal frequency range in the FFT power spectrum could reach at most the 11 kHz that corresponds to the Nyquist frequency. At 9.8 kHz we see that the signal compressed to 48 kbps strongly diverges away from the others, while at 11.5 kHz and above we observe the divergence of the signal compressed to 64 kbps.

The other three signals compressed at a higher rate (160, 128 and 96 kbps) maintain their parallelism throughout the whole of the 0–12 kHz frequency range. This parallelism is almost perfect because the differences in dB with respect to the original (averages: -1.38, -1.32 and -1.29 dB respectively for the three bitrates) are maintained with barely any variation in the whole of the frequency range. Pearson's correlation across the subjects between the mean original values and the mean compressed values, brings up the following values: 1 for the three highest bitrates, 0.998 for 64 kbps, 0.967 for 48 kbps and 0.947 for 32 kbps.

In order to better evaluate the correspondence between the original and compressed spectrum signals, an analysis of variance (ANOVA) 2×71 was carried out for the samples generated by each subject (in the case of 32 kbps compression, ANOVA was 2×65). The procedure was similar to that used by Mendoza *et al.* (16) in their study on gender differences in LTAS. Each analysis included a signal factor (S) (original signal vs. compressed signal) and a frequency level factor (F) along 71 (65) frequency levels. The results show that the relative correspondence between the original and compressed signal in the three highest compression rates is practically perfect (there is an absence of interaction between signal and frequency level). And that the small absolute difference of slightly more than 1 dB is maintained throughout the frequency levels and hence shows a significant main factor S. In fact, for 160 kbps, the main effects for the signal factor [$F(1,426) = 168.69$, $p < 0.001$] and level frequency factor [$F(70,426) =$

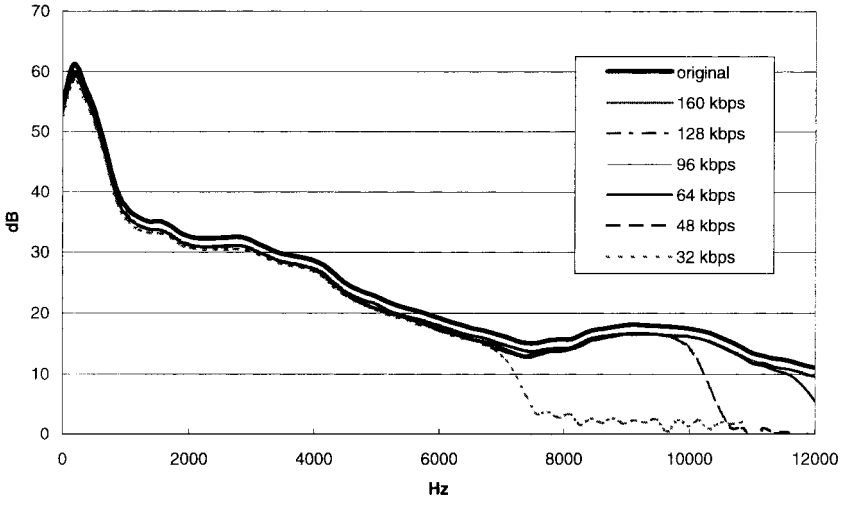

Fig. 1. LTAS of original and compressed voice signals. Representation of the mean values of amplitude (in decibels) in each frequency level analyzed in the range 0–12000 Hz (3).

98.63, $p < 0.001$] were significant. This was not the case for the interaction of signal × frequency [$F(70,426) = 0.01$, $p = 1$], which presented a very small F; this is indicative of a very strong parallelism between both signals for the whole of the frequency range studied (0–12 kHz). The pattern of results for 128, 96 and even 64 kbps is quite similar. For 128 kbps, the signal factor [$F(1,426) = 154.54$, $p < 0.001$] and level frequency factor [$F(70,426) = 97.41$, $p < 0.001$] were significant, but this was not the case for the interaction of signal × frequency level [$F(70,426) = 0.01$, $p = 1$]. For 96 kbps, the signal factor [$F = (1,426) = 147.77$, $p < 0.001$] and level frequency factor [$F(70,426) = 97.93$, $p < 0.001$] were significant but, once again, this was not the case for the interaction of signal × frequency level [$F(70,426) = 0.01$, $p = 1$]. For 64 kbps, the signal factor [$F(1,426) = 256.80$, $p < 0.001$] and level frequency factor [$F(70,426) = 96.91$, $p < 0.001$] were significant but this was not so for the interaction of signal × frequency level [$F(70,426) = 0.53$, $p = 0.999$]. The results were nevertheless of a distinct nature for the two lowest compression rates where an interaction of factors was obtained. Thus, for 48 kbps the signal factor [$F(1,426) = 868.52$, $p < 0.001$] and level frequency factor [$F(70,426) = 113.99$, $p < 0.001$] were significant; and the interaction signal × frequency level [$F(70,426) = 16.20$, $p < 0.001$] was also significant. For 32 kbps, the signal factor [$F(1,390) = 2768.51$, $p < 0.001$], level frequency factor [$F(64,390) = 132.78$, $p < 0.001$] and the signal × frequency level interaction [$F(64,390) = 40.33$, $p < 0.001$] were significant.

Discussion

The results show that for the whole of the frequency range 0–12 kHz which includes the most frequencies perceptually relevant to the speech signal, the compression produced by the MPEG-1 Layer 3 coding schema at the higher bitrates does not give rise to substantial changes in the signal. The LTAS of the signals compressed to 160, 128 and 96 kbps show a reduction of just over one decibel in its dynamic range but the relative distribution of energy throughout the frequency values is essentially the same as the original signal. The most important changes introduced by the compression algorithm occur for frequency values greater than 12 kHz and are of no consequence from a perceptual point of view yet they allow a substantial amount of bits to be saved. This close parallelism in the spectral profiles within the 0–12 kHz range is supported by a perfect correlation between the original and compressed signals. This parallelism is also reinforced by the absence of signal factor × frequency level factor interaction in the analysis of variance across the subjects. Hence, from the spectral composition point of view, the voice signal compressed to 160, 128 or 96 kbps remains unaltered in that frequency range at the time that we obtain reduction rates of 4.4:1, 5.5:1, and 7.4:1 respectively in relation to the original speech signal recorded at 44.1 kHz.

The compression performed at lower bitrates gives rise to changes that should be pointed out: at 64 kbps (compression rate of 11:1) the compressed signal spectrum diverges from the original beyond 11500 Hz, having maintained itself parallel up to this frequency. At a compression of 48 kbps (14.7:1), the spectral energy sharply diminishes beyond 9800 Hz, whilst at a compression of 32 kbps (22.1:1) this occurs at as low as 6800 Hz. At these lower bitrates hence—specially in the last two—high compression rates are attained but paying the price of causing major changes to the signal within the frequency range relevant to speech.

EXPERIMENT 2

The goal of the second experiment is to study how MPEG-1 Layer 3 compression at different bitrates affects a set of 29 acoustical voice parameters obtained from the Multi-Dimensional Voice Program (MDVP) model of Kay Elemetrics Corp. This program has become an important analytical tool that is increasingly used in voice studies, in the clinical setting as well as in research (4, 5, 13). This set of parameters includes those of a long-standing genre such as frequency and amplitude perturbation parameters (jitter, shimmer, etc.), noise to harmonics ratio, together with those more recently developed.

Method

Speakers. Subjects were 34 native speakers of *Valenciano* (Catalan) of both sexes (23 females and 11 males), students at the University Jaume I of Castellon (Spain), with ages ranging from 20 to 32. Each exhibited normal speech and audition.

Apparatus. The recording of voice samples was performed with a Shure SM58 microphone at a distance of about 15 cm from the mouth. The voice parameters were extracted with the MDVP model 4305 of Kay Elemetrics Corp.

Voice samples. Following the MDVP operations manual, the speakers were asked to produce a sustained phonation of /a/ vowel during 3 seconds at a comfortable pitch and loudness. The subjects were instructed to maintain as steady a phonation as possi-

ble. MDVP software can only work with two sample frequencies: 25 or 50 kHz. In this experiment, all voice samples were recorded at 50 kHz and directly stored in the host computer. The samples were recorded in a soundproof room at the University Laboratories.

Each voice sample was compressed by means of the Fraunhofer-Thomson compression scheme, which is the original and highest quality MPEG-1 Layer 3 algorithm available. The compression procedure involved a first step whereby the format of voice files was converted: the NSP file format by Kay Elemetrics Corp. was converted to WAV standard format, by way of the program *Sound File Converter* v. 3.1.0 by Bob Tice and Tom Carrell. This step involved a mere change of header of the voice file where extra non-audio information is included but does not affect the signal data. Once in WAV format, we applied the compression schema implemented in the *Cool Edit 2000* program by Syntrillium Software Corp.; the compressed voice file was then restored to NSP format for its acoustic analysis.

The compression algorithm does not allow that each bitrate can be associated with any output sample frequency in the codifying process; rather each bitrate is limited to specific sample frequencies to get a good result. Bearing in mind that the original signal was recorded at 50 kHz, all output sample frequencies were chosen in accordance with the recommendations derived from the compression algorithm in order to obtain the maximum sound quality within each bitrate. The only exception took place with the 48 kbps bitrate, which had a recommended optimal sample frequency of 32 kHz, but nevertheless 44.1 kHz was chosen to maintain the same conditions for the other bitrates. At the 32 kbps bitrate, the only output frequencies available were 24000 or 22050 Hz and the latter value was chosen following the recommendation. Thus, compressions were made at the following bitrates and sample frequencies: 160 kbps (44.1 kHz), 128 kbps (44.1 kHz), 96 kbps (44.1 kHz), 64 kbps (44.1 kHz), 48 kbps (44.1 kHz), and 32 kbps (22050 Hz). These values give a set of compression rates ranging from 5:1 to 25:1.

Given that all the compression options imply downsampling, an additional condition was stipulated for comparative purposes: i.e. that sample frequency of the original signal is converted from 50 to 44.1 kHz without there being any MPEG compression.

Acoustic analysis. All original and compressed voice samples were analyzed with the MDVP software and the following parameters were obtained:

Fundamental frequency parameters: Average Fundamental Frequency (Fo), Highest Fundamental Frequency (Fhi), Lowest Fundamental Frequency (Flo), Standard Deviation of Fo (STD), and Phonatory Fo-Range in semi-tones (PFR) for all extracted pitch periods.

Frequency perturbation parameters: *Absolute Jitter* (*Jita*) /μs/: It gives an evaluation in microseconds (μs) of the period-to-period variability of the pitch period within the analyzed voice sample. This measure is widely used in voice research (12) and is very sensitive to the pitch variations occurring between consecutive pitch periods. However pitch extraction errors may affect this measure, which is why it is of special interest regarding compression effects. *Jitter Percent* (*Jitt*) /%/: Relative period-to-period variability of the pitch period. *Relative Average Perturbation* (*RAP*) /%/: Introduced by Koike (14), this parameter gives the relative evaluation of the period-to-period variability of the pitch with smoothing factor of 3 periods. *Pitch Perturbation Quotient* (*PPQ*) /%/: Introduced by Koike *et al*. (15), it gives the variability of the pitch period at smoothing factor of 5 periods. *Smoothed Pitch Perturbation Quotient* (*sPPQ*) /%/: An evaluation of the long-term variability of the pitch period within the analyzed voice sample, with smoothing factor of 55 periods. RAP, PPQ and sPPQ have been extensively used in the last decade, given that they are less sensitive to pitch extraction errors due to smoothing in their calculation. *Fundamental Frequency Variation* (*vFo*) /%/: The relative standard deviation of the fundamental frequency. It reflects the very long-term variation of Fo within the analyzed voice sample. Any variations in the fundamental frequency are reflected in vFo, and this parameter increases regardless of the type of pitch variation, whether it be of the random or regular fluctuating type.

Amplitude perturbation parameters: *Shimmer in dB* (*ShdB*) /dB/: Evaluation in dB of the period-to-period variability of the peak-to-peak amplitude within the analyzed voice sample. As in other parameters, voice break areas are excluded. As occurs with jitter, this parameter has been widely used in voice research. *Shimmer Percent* (*Shim*) /%/: Relative evaluation of the period-to-period variability of the peak-to-peak amplitude. *Amplitude Perturbation Quotient* (*APQ*) /%/: Introduced by Koike *et al*. (15), it gives the relative evaluation of the variability of the peak-to-peak amplitude at smoothing of 11 periods. The smoothing reduces the sensitivity of APQ to pitch extraction errors. *Smoothed Amplitude Perturbation Quotient* (*sAPQ*) /%/: Evaluation of the long-term period-to-period variability of the peak-to-peak amplitude at smoothing of 55 periods. *Peak-Ampli-*

tude Variation (*vAm*) /%/: It gives the relative standard deviation of period-to-period calculated peak-to-peak amplitude. It reflects the very long-term amplitude variations within the analyzed voice sample.

Noise parameters: *Noise to Harmonic Ratio* (*NHR*): A general evaluation of the noise presence in the analyzed signal (such as amplitude and frequency variations, turbulence noise, subharmonic components or voice breaks). It is the ratio of inharmonic energy in the range 1500–4500 Hz to the harmonic spectral energy in the range 70–4500 Hz. *Voice Turbulence Index* (*VTI*): Ratio of the inharmonic energy in the range 2800–5800 Hz to the harmonic spectral energy in the range 70–4500 Hz. This parameter measures the relative energy level of high frequency noise, being a new attempt to compute breathiness in the voice signal. *Soft Phonation Index* (*SPI*): Ratio of the harmonic energy in the range 70–1600 Hz to the harmonic energy in the range 1600–4500 Hz. It is very sensitive to the vowel formant structure. This parameter is not actually a measurement of noise, but its formula is similar to the above two parameters and is therefore, as in the MDVP manual, listed in the same category.

Tremor parameters: *Fo-Tremor Frequency* (*Fftr*) /*Hz*/: It shows the frequency of the most intensive low frequency Fo-modulating component in the tremor range. *Amplitude Tremor Frequency* (*Fatr*) /*Hz*/: It shows the frequency of the most intensive low frequency amplitude modulating component in the tremor range. *Fo-Tremor Intensity Index* (*FTRI*) /%/: Ratio of the frequency magnitude of the most intensive low frequency modulating component (Fo tremor) to the total frequency magnitude of the analyzed signal. *Amplitude Tremor Intensity Index* (*ATRI*) /%/: Ratio of the amplitude of the most intensive low-frequency amplitude modulating component (amplitude tremor) to the total amplitude of the analyzed signal.

Parameters of Subharmonic components: *Number of Subharmonic Segments* (*NSH*): Number of subharmonic segments found during analysis. *Degree of Subharmonics* (*DSH*) /%/: Relative evaluation of subharmonic to Fo components in the analyzed sample.

Parameters of Voice irregularities: *Number of Unvoiced Segments* (*NUV*): Number of unvoiced segments detected during the analysis. *Degree of Voiceless* (*DUV*) /%/: Relative evaluation of non-harmonic areas in the voice sample.

Voice Break parameters: *Number of Voice Breaks* (*NVB*): Number of times the Fo was interrupted in the analyzed sample. *Degree of Voice Breaks* (*DVB*) /%/: Ratio of the length of areas representing voice breaks to the total sample length.

Results

As a first approach, in each compression condition (including the only downsampled signal condition), a multivariate analysis of variance (MANOVA) was performed considering parameter values as dependent variables and original vs compressed signal as factor. Parameters of voice break (NVB and DVB) were not included in the analysis because all values were nil in all conditions. The MANOVA (see Table 1) showed significant effects due to compression of voice signal only at bitrates of 48 kbps and 32 kbps. These results are congruent with the long-term average spectra obtained in the experiment 1, where the two lower bitrates yielded the profiles more deviated from the original. Data also show the relevance of the output sample frequency recommended by the algorithm Fraunhofer-Thomson: compression at 48 kbps, with a not recommended sample frequency, yield worse results [$F(1,33) = 65.97$, $p < 0.001$] than compression at 32 kbps with a recommended sample frequency [$F(1,33) = 12.86$, $p < 0.01$].

In order to study the extent to which the voice parameters of the compressed samples differ from those of the original samples, a discriminant analysis (DA) was conducted across subjects between the original signals and each compression condition; this analysis made use of voice parameters as discriminant variables. The greater the difference in the voice parameters between the original signals and the compressed ones, the greater the efficiency of the latter as variables of prediction in the classification of each sample. The results of this analysis are shown in Table 2. We can see that the percentage of correct classification between original samples and compressed ones at 160 kbps (54.8%) is very close to the chance level and is exactly the same when the sample is downsampled from 50 to 44.1 kHz. This data would indicate that by taking the voice parameters as predictable factors, both signals are practically indistinguishable from the original. When the signal is compressed to 128 or 96 kbps, the percentage of correct classification increases to 59.7%; this figure remains quite close to the random level. Below these bitrates, the discrimination between original and compressed signals increases noticeably, reaching its maximum at 48 kbps with 82.8% of correct classification, being the condition with a sample frequency not recommended by the compression algorithm. When considering variables that have a greater influence on the discriminant function, we see the emergence of the parameters concerned with the measurement of tremor, amplitude perturbation, noise, subharmonics and voice irregularities.

Although the MANOVA shows an overall non-significant result in a experimental condition, it is important to analyze the measures individually since some parameter could be severely affected. According to MANOVA and DA results, we hypothesize that the main discrepancies will have place in the lowest bitrates, especially at 48 and 32 kbps. Tables 3–6 show the parameters obtained from original and compressions at different bitrates, classified in four groups. Furthermore, the tables include the original signal after it was simply downsampled from 50 to 44.1 kHz (hereafter referred to as downsampled signal). In the head of each compression condition, bitrate, sampling rate and compression ratio are indicated. In the first column of data means and standard deviations across 34 subjects of the MDVP parameters obtained from the original voice signal are shown. Each box of the compression conditions shows the mean and standard deviation of the percentage differences and the Pearson correlation between original and compressed signal. Percentage difference of each parameter was calculated this way: ((original-compressed)/original) $\times$ 100. To separate compression from merely downsampling effects the following operations have been made: we calculated the percentage difference between the original and all other conditions for every parameter; then we compared (t-test) the percentage of difference between original and downsampled (*ds*) to the differences between original and all compression bitrates (*do*). Any significant differences in the t-test at the 0.05 or 0.01 levels are indicated in the tables.

Table 3 presents the results obtained for the Fundamental Frequency and Frequency Perturbation parameters. Generally speaking, very high correlations are observed: these are greater for medium and higher bitrates—64 kbps or more—and in the downsampled signal. Fo is maintained almost exactly the same as the original value for all compression conditions (unless otherwise stated, the downsampled signal value is included under this general term) yielding perfect correlations. The miniscule differences obtained in the order of thousandths of Hz are in some cases significant as they are generated in a systematic manner across the voice samples. Thus, a mean difference of 0.017 Hz between Fo of the original signal and the compressed at 160 kbps (170.348–170.331 Hz) is significant because it is not derived from the random differences across the subjects, rather in 33 of the 34 subjects the Fo of the compressed signal is a few thousandths of Hz less than original. When we compare the percent difference original—160 kbps to the percent difference original—downsampled the discrepancy is inferior to 0.01%, but it is significant

Table 1. *Results of MANOVA in each condition with acoustical voice parameters obtained in MDVP as dependent variable and original vs compressed (downsampled) signal as factor. In each compression condition, bitrate, sampling rate and compression ratio are indicated*

Downsampled 44100 (1.3:1)	MPEG 160 kbps 44100 (5:1)	MPEG 128 kbps 44100 (6.3:1)	MPEG 96 kbps 44100 (8.3:1)	MPEG 64 kbps 44100 (12.5:1)	MPEG 48 kbps 44100 (16.7:1)	MPEG 32 kbps 22050 (25:1)
$F(1,33) = 0.46$ $p = 0.502$	$F(1,33) = 2.42$ $p = 0.129$	$F(1,33) = 3.29$ $p = 0.079$	$F(1,33) = 2.57$ $p = 0.118$	$F(1,33) = 1.25$ $p = 0.273$	$F(1,33) = 65.97$ $p = 0.000$	$F(1,33) = 12.67$ $p = 0.001$

[$t(33) = 2.33$, $p < 0.05$] because of its systemacity. The difference of as little as -0.02% in MPEG compression at 48 kbps is significant at the 0.01 level because in 18 subjects of the 34, the Fo value is a few hundredths of Hz greater than in the original signal.

The compressed-original discrepancies are somewhat bigger in the extremes of the Fo range exhibited by each sample. These discrepancies are more pronounced in the upper extreme (Fhi) than in the lower extreme (Flo): percent differences in Fhi of up to -2.01% (at 48 kbps) and differences in Flo of up to 1.11% (at 48 kbps) with respect to the original value. Generally speaking, there is a loss of fidelity over the frequency range starting at bitrates equal to or less than 64 kbps. The PFR does not deviate from the original value by more than 4.90% in the 96–160 bitrate range while there is a percent divergence of 11.86%, 21.67%, and 18.97% for compressions at 64, 48 and 32 kbps respectively. A similar phenomenon occurs with the STD. On the other hand, there are poorer results generated from the bitrate at 48 kbps than from the bitrate at 32 kbps; this occurs in the majority of parameters. As mentioned previously, this is probably due to a non-optimal combination of sample rate and bitrate. We must bear in mind that for the compression set at 48 kbps, the chosen sample rate was 44.1 kHz as we wanted to level up with the rest of the higher-value bitrates. Nevertheless, the preferred sample rate for 48 kbps according to the Frauenhofer-Thomson compression scheme is not 44.1 kHz but 32 kHz.

Absolute jitter measurements (Jitta) maintain their fidelity in all conditions except for compressions at 48 kbps. In the rest of conditions the discrepancies were small and not significant, being the maximum difference 1.206 μs for the compression at 96 kbps. The compression at 48 kbps is a case in itself as it generated a significant difference of -15.449 μs and the mean of percent differences is -22.92% from original. The relative jitter (Jitt) generates results that are parallel to absolute values. The remaining frequency perturbation parameters, RAP, PPQ, sPPQ and vFo, show a general pattern of similar results: i.e. the poorest results for the bitrate at 48 kbps and, at a great distance, results for bitrate at 32 kbps. Within the range of medium and higher bitrates (64–160 kbps), compression at 64 kbps generated the worst results without there being any notable differences among the remaining bitrates. The 96–160 kbps compressions closely approximated original values with maximum discrepancies only of 1.43% in relation to the original. We can also argue that, generally speaking, the downsampled signal shows no marked difference with respect to the high bitrates fitting to their

Table 2. *Discriminant analysis between original and compressed samples utilizing MDVP parameters as discriminating variables*[a]. *The six highest correlations (in absolute values) between variables and discrimination function are presented. Last row shows percentage of correct classification*

		MPEG 160 kbps	MPEG 128 kbps	MPEG 96 kbps	MPEG 64 kbps	MPEG 48 kbps	MPEG 32 kbps
	44100	44100	44100	44100	44100	44100	22050
	(1.3:1)	(5:1)	(6.3:1)	(8.3:1)	(12.5:1)	(16.7:1)	(25:1)
Correlations variables—discrimination function	Fatr (−0.348)	Fftr (0.213)	SPI (0.276)	SPI (−0.419)	Fatr (0.273)	sAPQ (−0.481)	Shim (0.367)
	DSH (0.268)	Fatr (0.172)	Fatr (0.267)	VTI (0.206)	ATRI (0.263)	APQ (−0.336)	ShdB (0.352)
	NSH (0.254)	vAm (−0.157)	Fftr (0.173)	Fatr (−0.176)	APQ (0.222)	ShdB (−0.335)	APQ (0.342)
	VTI (−0.147)	NHR (−0.139)	VTI (−0.167)	APQ (0.098)	Shim (0.217)	Shim (−0.334)	Fatr (0.308)
	Fftr (−0.117)	PFR (−0.101)	PFR (0.165)	vFo (−0.097)	DUV (0.206)	Fatr (−0.331)	PFR (0.179)
	vAm (0.109)	vFo (−0.089)	vAm (−0.127)	FTRI (−0.095)	NUV (0.205)	NUV (−0.285)	ATRI (0.168)
Correct classification	54.8%	54.8%	59.7%	59.7%	69.9%	82.8%	73.0%

[a]Parameter abbreviations are explained in the text.

Table 3. *Fundamental frequency and frequency perturbation parameters obtained in MDVP. In the first column of data means and standard deviations (between parenthesis) across 34 subjects of values from original signal are expressed. In compression conditions (included downsampled signal) means and standard deviations (between parentheses) of the percentage differences and Pearson correlation between original and compressed signal are expressed. Correlations below 0.90 are written in bold type. In the head of each compression condition, bitrate, sampling rate and compression ratio are indicated.*

	ORIG. 50000	44100 (1.3:1)	MPEG 160 kbps 44100 (5:1)	MPEG 128 kbps 44100 (6.3:1)	MPEG 96 kbps 44100 (8.3:1)	MPEG 64 kbps 44100 (12.5:1)	MPEG 48 kbps 44100 (16.7:1)	MPEG 32 kbps 22050 (25:1)
Fo (Hz)	170.348	0.00	0.00 *	0.00	0.00	0.00	−0.02 **	0.00
	(43.757)	(0.01) $r = 1$	(0.01) $r = 1$	(0.01) $r = 1$	(0.01) $r = 1$	(0.01) $r = 1$	(0.03) $r = 1$	(0.02) $r = 1$
Fhi (Hz)	183.809	0.07 °	0.36	−0.27	0.64	−0.12	−2.01 **	−1.01
	(51.147)	(0.20) $r = 1$	(2.04) $r = 0.993$	(1.41) $r = 0.998$	(3.39) $r = 1$	(3.42) $r = 1$	(3.65) $r = 0.990$	(2.64) $r = 0.993$
Flo (Hz)	160.485	0.05	0.00	0.05	−0.15	0.60	1.11 *	1.07
	(41.575)	(0.26) $r = 1$	(0.10) $r = 1$	(0.26) $r = 0.999$	(2.39) $r = 0.995$	(2.80) $r = 0.993$	(2.43) $r = 0.992$	(3.14) $r = 0.988$
STD (Hz)	2.336	−0.16	0.85	0.47	1.42	−4.07	−21.10 **	−9.82 **
	(1.288)	(2.98) $r = 0.993$	(2.45) $r = 0.995$	(3.29) $r = 0.994$	(4.87) $r = 0.990$	(10.90) $r = 0.981$	(26.23) $r = 0.941$	(19.29) $r = 0.951$
PFR	3.235	0.49	1.18	−4.90	−1.97	−11.86	−21.67 **	−18.97 **
(semitone)	(1.724)	(2.86) $r = 0.995$	(6.86) $r = 0.980$	(19.04) $r = 0.972$	(23.37) $r = \mathbf{0.876}$	(36.23) $r = \mathbf{0.841}$	(31.91) $r = \mathbf{0.896}$	(36.20) $r = \mathbf{0.881}$
Jita (us)	64.011	−0.08	0.97	−0.09	0.91	−2.67	−22.92 **	−1.12
	(39.083)	(3.74) $r = 0.995$	(3.08) $r = 0.997$	(3.81) $r = 0.996$	(4.69) $r = 0.997$	(9.64) $r = 0.991$	(24.75) $r = 0.946$	(17.92) $r = 0.978$
Jitt (%)	1.051	−0.09	0.98	−0.09	0.91	−2.68	−22.93 **	−1.12
	(0.669)	(3.74) $r = 0.995$	(3.08) $r = 0.997$	(3.81) $r = 0.996$	(4.69) $r = 0.997$	(9.62) $r = 0.990$	(24.75) $r = 0.951$	(17.93) $r = 0.981$
RAP (%)	0.629	0.21	0.96	0.19	1.0	−2.00	−22.25 **	−0.36
	(0.411)	(3.70) $r = 0.996$	(3.30) $r = 0.997$	(3.85) $r = 0.996$	(5.45) $r = 0.996$	(10.12) $r = 0.991$	(25.38) $r = 0.953$	(18.66) $r = 0.981$
PPQ (%)	0.615	−0.23	1.23 *	−0.04	0.98	−3.03	−25.13 **	−2.55
	(0.393)	(3.77) $r = 0.995$	(3.03) $r = 0.997$	(3.82) $r = 0.995$	(4.05) $r = 0.996$	(9.81) $r = 0.989$	(25.92) $r = 0.947$	(18.33) $r = 0.980$
SPPQ (%)	0.799	−0.33	0.44	−0.30	0.93 *	−1.31	−13.22 **	−1.56
	(0.345)	(3.41) $r = 0.994$	(2.06) r = 0.998	(3.48) $r = 0.994$	(2.71) $r = 0.997$	(4.90) $r = 0.994$	(15.33) $r = 0.950$	(9.70) $r = 0.984$
Vfo (%)	1.351	−0.16	0.85	−0.47	1.43	−4.06	−21.07 **	−9.82 **
	(0.599)	(2.97) $r = 0.991$	(2.44) $r = 0.993$	(3.28) $r = 0.992$	(4.86) $r = 0.987$	(10.90) $r = 0.971$	(26.20) $r = 0.923$	(19.28) $r = 0.943$

Parameter abbreviations are explained in the text.
do = percentage difference between the original and each compression bitrate.
ds = percentage difference between the original and the downsampled (at 44100 Hz) signal.
°: mean difference between original and downsampled signal at a significance value $p < 0.05$ (*t*-test).
*: mean difference between *do* and *ds* at a significance value $p < 0.05$ (*t*-test).
**: mean difference between *do* and *ds* at a significance value $p < 0.01$ (*t*-test).

original values in both the frequency fundamental and the frequency perturbation parameters.

With regard to the amplitude perturbation parameters (Table 4) fidelity is very high for the downsampled signal and for signals compressed to 128 and 160 kbps. In these signals, no discrepancy attains 1% with respect to the original in any parameter. The compression at 96 kbps presents somewhat greater discrepancies but still at overall low levels: no parameter diverges by more than 2.17% from the original. The results generated below this bitrate are much poorer and especially so in the shorter term parameters (ShdB, Shim, APQ). On the other hand, the vAm is more resistant to modification. MPEG 64 kbps compression diverges by 15% in the first three parameters (ShdB, Shim, and APQ), by 5.52% in sAPQ and by 0.21% in vAm. For lower bitrates, the signal undergoes significant change. MPEG 48 kbps diverges about 80% in the ShdB, Shim and APQ parameters (with compressed-original correlations below 0.90), by 35.64% in sAPQ (correlation $r = 0.888$) and by 6.79% in vAm. MPEG 32 kbps diverges by 35% in the two Shimmer cases, 39.56% in the APQ, 19.93% in sAPQ and by 4.58% in vAm.

Noise parameters (Table 5) are of special interest as they seem to be particularly sensitive to any possible noise introduced through the compression system. We nevertheless found values in consonance with the original for the MPEG compressions. The averages of the NHR and the VTI practically coincide with the original signal for both the downsampled signal and the 96–160 kbps compressions. The correlation is almost perfect for NHR and decreases slightly in the VTI. MPEG 64 kbps values diverge an average of 1.77% of the original NHRs and the correlation with the original VTIs decreases to 0.836. MPEG at 48 kbps diverges from the original by 13.33% and 10.26% for both parameters respectively. MPEG 32 kbps diverges by 3.56% in NHR and the VTI correlation decreases to 0.769. The SPI is not strictly speaking a noise parameter, even though it is included in the MDVP manual in this section due to the similarity of the calculation; its values are maintained quite close to the original in all the bitrates, including the lowest ones. The maximum difference is 1.65% at 32 kbps and all the correlations are almost perfect. In fact, these small differences become significant due to their systematic occurrence in the voice samples. This fidelity shows that, even in the most intensive compressions, the spectral form of the signal is maintained regarding the ratio between harmonic energies for values both below and above 1600 Hz, within the 0–4500 Hz range.

Tremor parameters (Table 6) appear to be the most sensitive to MPEG compression of the signal, spe-

Table 4. *Amplitude perturbation parameters obtained in MDVP. See title of Table 3*

	ORIG. 50000	44100 (1.3:1)	MPEG 160 kbps 44100 (5:1)	MPEG 128 kbps 44100 (6.3:1)	MPEG 96 kbps 44100 (8.3:1)	MPEG 64 kbps 44100 (12.5:1)	MPEG 48 kbps 44100 (16.7:1)	MPEG 32 kbps 22050 (25:1)
ShdB (dB)	0.369	−0.07	0.30	−0.01	−1.96 **	−14.57 **	−79.09 **	−34.68 **
	(0.161)	(1.05) $r = 0.999$	(0.78) $r = 0.999$	(1.09) $r = 0.999$	(2.94) $r = 0.999$	(17.45) $r = 0.944$	(74.96) $r = \mathbf{0.793}$	(42.75) $r = 0.937$
Shim (%)	4.211	−0.04	0.31	0.01	−2.17 **	−15.59 **	−81.06 **	−35.42 **
	(1.828)	(0.98) $r = 0.999$	(0.87) $r = 0.999$	(1.07) $r = 0.999$	(2.93) $r = 0.999$	(17.89) $r = 0.939$	(74.55) $r = \mathbf{0.780}$	(41.94) $r = 0.934$
APQ (%)	3.157	−0.11	0.16	−0.23	−2.12 **	−14.88 **	−77.83 **	−39.56 **
	(1.286)	(0.99) $r = 0.999$	(0.81) $r = 0.999$	(1.05) $r = 0.999$	(2.17) $r = 0.999$	(14.33) $r = 0.930$	(52.45) $r = \mathbf{0.823}$	(32.19) $r = 0.933$
sAPQ (%)	5.052	−0.16	−0.10	−0.22	−1.13 **	−5.52 **	−35.64 **	−19.93 **
	(1.547)	(0.76) $r = 0.999$	(0.61) $r = 0.999$	(0.80) $r = 0.999$	(1.41) $r = 0.999$	(8.37) $r = 0.963$	(25.56) $r = \mathbf{0.888}$	(16.91) $r = 0.950$
vAm (%)	16.221	−0.15	−0.55	−0.21	−0.22	0.21	−6.79 **	−4.58 **
	(6.192)	(1.47) $r = 0.999$	(2.11) $r = 0.998$	(1.47) $r = 0.999$	(2.52) $r = 0.998$	(9.63) $r = 0.953$	(11.71) $r = 0.982$	(7.83) $r = 0.992$

Parameter abbreviations are explained in the text.

cially Fo – Tremor Frequency (Fftr) and Amplitude Tremor Frequency (Fatr). Fftr presents discrepancies with respect to the original of 7–12% in the 96–160 kbps bitrates, giving correlations lower than 0.90. The divergence is greatest in the other bitrates, with correlations as little as $r = 0.499$ at 64 kbps, $r = 0.448$ at 48 kbps, and $r = 0.788$ at 32 kbps. Fatr attains differences of up to 8% in the high bitrates while there is also a marked deterioration in the bitrates equal to or below 64 kbps. In MPEG 48 kbps, the correlation between the Fatr of the original signals and the compressed ones is nil ($r = 0.005$). When we examine the data for each sample, we can see that in some cases the Fatr varies, for example, from 2.2, 1.9 or 1.5 Hz in the original samples to values such as 9.5, 14.8 or 16.7 Hz in the compressed ones. Furthermore, the simple downsampling of the signal introduces a difference of 1.70% in Fftr and 4.96% in Fatr. FTRI and ATRI show a greater degree of stability, even though there is some deterioration in the lower bitrates.

In relation to parameters of subharmonic components, we have to stress that statistics for central tendency are not useful as they are nil for most of the values. From a total of 34 original samples, only 6 presented any subharmonic component (to be precise, 5 samples with 1–3 subharmonics and one sample with 15 subharmonics). Both the downsampled signals as well as the compressed signals at 64–160 kbps reproduce these calculations with a high degree of precision; there is an oscillation of only ± 1 subharmonic in some subjects. At MPEG 48 kbps there are 7 subjects with some harmonics that emerge from the compressed sample whereas the original sample does not present any. In MPEG 32 kbps, a sample loses its single subharmonic and the quantity of subharmonics in the rest of the samples varies considerably from the original. These data are reflected in NSH as well as DSH.

Of all the subjects studied, only one presented a voice irregularity—unvoiced segment—in the original signal (NUV = 3, DUV = 3.125). These parameters are reproduced in the downsampled signal as well as in the higher bitrates (96–160 kbps). In MPEG 64 kbps there is a compressed sample with a voice irregularity that does not exist in the original. This also occurs in MPEG 48 kbps and MPEG 32 kbps, eight and six times respectively.

Finally, given the nature of the recorded signal, a sustained phonation of /a/, there is not a single voice break in any of the original samples (NVB = 0 and DVB = 0), and this is maintained in all the compressions.

Table 5. *Noise parameters obtained in MDVP. See title of Table 3*

	ORIG. 50000	44100 (1.3:1)	MPEG 160 kbps 44100 (5:1)	MPEG 128 kbps 44100 (6.3:1)	MPEG 96 kbps 44100 (8.3:1)	MPEG 64 kbps 44100 (12.5:1)	MPEG 48 kbps 44100 (16.7:1)	MPEG 32 kbps 22050 (25:1)
NHR	0.135	0.06	0.10	0.08	0.08	−1.77 **	−13.33 **	−3.56 **
	(0.020)	(0.26) $r = 0.999$	(0.30) $r = 0.999$	(0.28) $r = 0.999$	(0.49) $r = 0.999$	(1.29) $r = 0.996$	(8.50) $r = \mathbf{0.863}$	(2.27) $r = 0.987$
VTI	0.053	−2.05	−1.57	−1.46	−1.72	−2.26	−10.26 *	−6.94
	(0.017)	(16.43) $r = 0.923$	(12.86) $r = 0.949$	(16.39) $r = 0.934$	(17.13) $r = 0.901$	(20.41) $r = \mathbf{0.836}$	(29.46) $r = \mathbf{0.713}$	(25.53) $r = \mathbf{0.769}$
SPI	9.071	−0.56	−0.56	−0.74 **	−.93 **	−1.05 **	−0.98	−1.65 **
	(4.215)	(0.58) $r = 0.999$	(0.51) $r = 0.999$	(0.50) $r = 1$	(0.80) $r = 0.999$	(0.77) $r = 0.999$	(1.70) $r = 0.999$	(1.60) $r = 0.999$

Parameter abbreviations are explained in the text.

Discussion

The results of this experiment show that for high compression bitrates such as 160, 128 and even 96 kbps, the voice parameters calculated from the compressed samples generally maintain fidelity to the parameters from the original samples. Nevertheless, in the lower bitrates of the study (32, 48 and in some cases 64 kbps) some of these parameters are greatly distorted. When the voice signal recorded at 50 kHz in the MDVP is compressed via the Fraunhofer-Thomson Scheme, a downsampling to 44.1 kHz also takes place; this is the highest optimal output frequency available for the higher bitrates. It must be stressed that the voice parameters obtained from 96 to 160 kbps compressed samples do not in fact vary more than when the signal is simply downsampled from 50 to 44.1 kHz, without there being any compression applied.

The compression schema introduces a very tiny variation of the fundamental frequency of voice in the order of a few hundredths of a hertz, which is irrelevant from a practical point of view, as the correlation is perfect with respect to the original. These modifications occur equally when the signal only is downsampled and are much lower than in the case when directly digitized samples are compared to taped voice samples (Gelfer and Fendel (7), found a variation of approximately 3.2 Hz and a correlation $r = 0.989$). The classic frequency perturbation parameters used in the speech clinic, such as jitter, whether it be measured in absolute or relative terms, or indeed the widely used RAP with a smoothing factor of three periods, as well as others with greater smoothing factors are quite close to those of the original signal. We found that for example absolute jitter from compressed samples does not differ on average by more than 1.3 μs from the value calculated from the original samples. If we bear in mind that the range 80–100 μs establishes the borderline region between normal and pathological voice production (9, 1) fluctuations of this magnitude do not present any clinical significance.

The amplitude perturbation parameters reflect a high degree of fidelity in the superior bitrates even though we are dealing with measurements that are highly sensitive to any manipulation of the signal. Contrary to what happens with jitter, Gelfer and Fendel (7) found that shimmer loses precision when taped voice samples are compared with directly digitized samples. The correlation between the shimmer values calculated in both recording procedures was very low ($r = 0.481$). Our data showed correlations of 0.999 in all the parameters comprising this class (ShdB, Shim, APQ, sAPQ and vAm), with averages

Table 6. *Tremor parameters obtained in MDVP. See title of Table 3*

	ORIG. 50000	44100 (1.3:1)	MPEG 160 kbps 44100 (5:1)	MPEG 128 kbps 44100 (6.3:1)	MPEG 96 kbps 44100 (8.3:1)	MPEG 64 kbps 44100 (12.5:1)	MPEG 48 kbps 44100 (16.7:1)	MPEG 32 kbps 22050 (25:1)
Fftr (Hz)	3.269	−1.70	−12.64	−11.30	−7.48	−33.04	−99.67	−6.87
	(3.353)	(6.26) $r = 0.998$	(64.98) $r =$ **0.882**	(64.12) $r = 0.929$	(37.13) $r =$ **0.801**	(116.09) $r =$ **0.499**	(365.79) $r =$ **0.448**	(39.48) $r =$ **0.788**
Fatr (Hz)	2.619	−4.96	−4.89	−7.98	3.08	−21.26	−103.48 *	−64.63
	(1.212)	(26.00) $r = 0.948$	(26.33) $r = 0.947$	(30.82) $r = 0.912$	(12.67) $r =$ **0.860**	(83.14) $r =$ **0.560**	(231.84) $r =$ **0.005**	(204.87) $r =$ **0.175**
FTRI (%)	0.386	1.29	−0.22	1.34	0.83	−1.86	−11.50 **	−5.81
	(0.146)	(4.75) $r = 0.987$	(5.95) $r = 0.982$	(5.06) $r = 0.985$	(9.30) $r = 0.954$	(22.12) $r = 0.948$	(25.39) $r =$ **0.874**	(19.95) $r = 0.906$
ATRI (%)	5.307	−0.65	−1.37	−3.21	2.95	−6.32	−40.54	−2816
	(2.541)	(3.77) $r = 0.998$	(4.18) $r = 0.998$	(12.76) $r = 0.977$	(12.18) $r = 0.978$	(25.85) $r = 0.915$	(119.38) $r =$ **0.558**	(86.17) $r =$ **0.711**

Parameter abbreviations are explained in the text.

that are very close to the original at bitrates of 160, 128 and to a lesser extent, 96 kbps. The percent shimmer (Shim), for example, varies by no more 0.31% on average in the first two bitrates. This variation is irrelevant from a clinical point of view if we consider that normative data would hold expected shimmer in normal voices at a maximum of 4%—see MDVP manual or Glaze *et al.* (8).

Contrary to what was first expected, noise parameters maintained a fairly high fidelity for the higher bitrates, especially with regard to the NHR and SPI. In both cases, the correlation between compressed and original samples was almost perfect. In the case of the VTI—given that it is a parameter that includes in its calculation noise of a higher frequency range (2800–5800 Hz) where variability is greater—we found slightly smaller correlations but in all cases greater than 0.90. These parameters show the relationship between the noise components and the harmonic components of the signal in different sections of the frequency range (NHR and VTI) or indeed between the harmonic components above and below 1600 Hz (SPI). This fidelity reveals that the compression code does not introduce any major changes in the spectrum of the signal, at least for the frequency range considered in the calculation of these parameters (0–4500 Hz for NHR and SPI; 0–5800 Hz for VTI). This is a breakthrough as noise parameters are quite sensitive to signal manipulation. In a recent work concerning the suitability of Minidisk (MD) recordings for voice perturbation analysis, Winholtz and Titze (17) concluded that no distortions were introduced by compression caused by the MD technique. The authors observed that not a single perturbation parameter underwent a major change except for the signal-to-noise ratio, which was approximately 10 dB less for MD recordings than normal DAT recordings. In accordance with our parameters, the MPEG compression at high bitrates gives a better signal/noise relationship.

The values of the parameters related to the subharmonic components and voice irregularity did not change when the signal is compressed to high bitrates. For low bitrates nevertheless, the algorithm of compression sometimes introduced subharmonic components or voice irregularities that were not present in the original signal.

Finally, tremor parameters seem to demonstrate a greater degree of sensitivity to signal compression. In fact, Fftr is the only parameter that presents a correlation lower than 0.90 in one of the two higher bitrates ($r = 0.882$ at 160 kbps). The deterioration of these parameters (Fftr, Fatr, FTRI, ATRI) is more pronounced in bitrates equal or inferior to 64 kbps, so much so that in some of them the correlations with the original values are very low or nil. The extraction process of the tremor parameters yields the amplitude and frequency demodulation curves of the voice signal. These curves contain information about the long-term amplitude and frequency variability of the voice signal. At low bitrates, the compression process mainly deteriorates the periodic sequence of amplitude and frequency values along the signal (Fatr and Fftr parameters). Working at low bitrates, if the encoder runs out of bits, it will not encode some blocks of signal data with the required fidelity (2). This loss of fidelity has consequences in the fine-grain structure of the sound wave, changing the tremor parameters.

In sum, the compressed voice signal in accordance with MPEG-1 Layer 3 Codec sustains a high degree of fidelity to the main voice parameters which are familiar in clinical research and practice and which furthermore have been extracted by virtue of MDVP software by Kay Elemetrics Corp. This fidelity exists so long as compression takes place at high bitrates (at around 160, 128 or even 96 kbps) which, all the same, generate high compression ratios (5:1, 6.3:1 and 8.3:1 respectively per channel) when the original signal is recorded at 50 kHz. A clear indicator of this fidelity in higher bitrates is the difficulty that discriminant analysis has in separating the original samples from the compressed ones, taking the voice parameters as variables of prediction. Lower bitrates such as 64, 48 or 32 kbps generate even higher compression figures, but the signal is substantially modified, resulting in major changes in most of the parameters studied. It is known that perceptual encoders when run at too low bitrates or with the wrong parameters show sound deficiencies by the error introduced in the compression process. These deficiencies consist of coding artifacts mainly causing a loss of bandwidth as consequence that some high frequency content is lost. The most common case is that the loss of bandwidth is not constant, but time varying for what the effect becomes more unsatisfactory (2). These alterations are less pronounced at the 64 kbps bitrate than in the other two lower ones. We should take note that the signal which suffers the greatest modification is the one compressed to 48 kbps; even though this bitrate is greater than 32 kbps, we should keep in mind that it includes a sample frequency which is not recommended as optimal as dictated by the compression algorithm.

Worthy of special interest in the study of compressed signals is the comparison of spectral composition to the original signal. Fidelity to the original values is maintained for noise parameters at high bitrates. Given that in the calculation of the noise parameters, there is a comparison of spectral energies

corresponding to different frequency ranges within 0–5800 Hz, the fidelity of these parameters, together with LTAS results from experiment 1, makes us think that the spectral composition of the compressed voice maintains its accuracy within this range.

CONCLUSIONS

The compression schema MPEG-1 Layer 3, or mp3, has had a great impact given that very big compression ratios are reached at the same time that it maintains high sound quality. Listening tests carried out with trained engineers and musicians have shown that when compression takes place at high bitrates, the signal, for both music and speech, is virtually indistinguishable from the original. However, we do not know the magnitude of change that such a compression system can introduce in the voice signal and to what extent it can alter parameters that are relevant in the investigation and clinical practice. In this study the Fraunhofer-Thomson compression scheme has been applied, being the original and highest quality algorithm available. The results obtained through the MDVP in a group of 29 parameters of great sensitivity to the manipulation of the signal, as well as the comparison of the spectra between compressed and original speech samples allow to state the following conclusions: a) The compression to high bitrates, of the order of 160 or 128 kbps, produce similar LTAS of the signal in the frequency range 0–12 kHz, and the modifications introduced in the voice parameters are minimal, so that all the correlations original-compressed signals are superior at 0.90, except for Fftr. These modifications are not greater than those that arise when the voice signal is simply downsampled from 50 to 44.1 kHz and they do not have any clinical significance by themselves to modify a clinical diagnosis or an evaluation pre-post treatment. b) In the high bitrates the most sensitive parameters to the compression are those related with the measure of tremor, Fo – Fftr and Fatr. c) The compression of the voice to same or inferior bitrates to 64 kbps alters the spectral composition of the signal significantly at the highest frequencies in the range 0–12 kHz and introduces major modifications in some of the studied parameters. The deterioration of the signal is greater as the bitrate of compression decreases, the most affected parameters being those related with the measures of tremor, amplitude perturbation, noise, subharmonics and voice irregularities. For this reason, it would be unadvisable in clinical practice to compress voices at bitrates equal to or below 64 kbps. d) The output sample frequency recommended by the Fraunhofer-Thomson algorithm has been shown to be of great importance since the compressed signal to 48 kbps with a not recommended frequency clearly obtains worse results than the compression to 32 kbps with a recommended frequency.

Nevertheless it is necessary to keep in mind that this test of MPEG compression effect has been performed with normal voice signals. Compression of pathological voices with a very degraded harmonic structure (e.g. esophageal speech) could yield quite different results from those obtained in normal voices. Before we use MPEG compression in the daily clinical practice it will be necessary that further research broadens the scope of this first study to pathological voices and applying other test methods.

ACKNOWLEDGEMENTS

This work was supported by *Fundació Caixa Castelló-Bancaixa* and the University Jaume I of Castellon, Spain, Project P1A99-01. The authors would like to thank the helpful comments of two anonymous reviewers on an earlier version of the manuscript. Translated by John Joseph Velez B.A. (University of Melbourne).

REFERENCES

1. Baken RJ, Orlikoff RF. Clinical measurement of speech and voice. San Diego, CA: Singular Thomson Learning, 2000.
2. Brandenbug K. MP3 and AAC explained. AES 17th International Conference on High Quality Audio Coding, 1999.
3. Brandenburg K, Stoll G. The ISO/MPEG-1 Audio Codec: A Generic Standard for Coding of High Quality Digital Audio. J. Audio Eng Soc 1994; 42: 780–92.
4. Cimino AM, Sapienza C. Reliability of the Multidimensional Voice Program (MDVP) for acoustic analysis. Meeting of The Voice Foundation, 1999.
5. Corina J, Hilgers FJM, Verdonck-de-Leeuw IM, Koopmans-van Beinum FJ. Acoustical analysis and perceptual evaluation of tracheoesophageal prosthetic voice. J Voice 1998; 12: 239–48.
6. EBU. Basic audio quality requeriments for digital audio bit-rate reduction systems for broadcast emission and primary distribution. CCIR document number TG 10-2/3, 1991;28.
7. Gelfer MP, Fendel DM. Comparisons of jitter, shimmer, and signal-to-noise ratio from directly digitized versus taped voice samples. J Voice 1995; 9: 378–82.
8. Glaze L, Bless D, Susser R. Acoustic analysis of vowel and loudness differences in children's voices. J Voice 1990; 4: 37–44.
9. Horii Y. Fundamental frequency perturbation observed in sustained phonation. J Speech Hear Res 1979; 22: 5–19.
10. ISO/IEC JTC 1/SC29/WG11 MPEG, International Standard IS 11172-3. Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mb/s, Part 3:Audio.

11. ITU-R. Methods for the subjective assessment of small impaiments in audio systems including multichannel sound systems. ITU-R Recommendation BS.1116. 1994. Geneva.
12. Iwata S, von Lenden H. Pitch perturbations in normal and pathological voices. Folia Phoniatr 1970; 22: 117–28.
13. Kent RD, Vorperian HK, Duffy JR. Reliability of the Multi-Dimensional Voice Program for the analysis of voice samples of subjects with dysarthria. Am J Speech-Lang Pathol 1999; 8: 129–36.
14. Koike Y. Application of some acoustic measures for the evaluation of laryngeal dysfunction. Studia Phonologica 1973; 7: 17–23.
15. Koike Y, Takahashi H, Calcatera T. Acoustic measures for detecting laryngeal pathology. Acta Otolaryngol 1977; 84: 105–17.
16. Mendoza E, Valencia N, Muñoz J, Trujillo H. Differences in voice quality between men and women: use of the Long-Term Average Spectrum (LTAS). J Voice 1996; 10: 59–66.
17. Winholtz WS, Titze IR. Suitability of minidisc (MD) recordings for voice perturbation analysis. J Voice 1998; 12: 138–42.