Acoustic Analysis of Pathological Voices Compressed with MPEG System

*Julio Gonzalez, †Teresa Cervera, and *M. José Llau

Castellon, Spain

Summary: The MPEG-1 Layer 3 compression schema of audio signal, commonly known as *mp3*, has caused a great impact in recent years as it has reached high compression rates while conserving a high sound quality. Music and speech samples compressed at high bitrates are perceptually indistinguishable from the original samples, but very little was known about how compression acoustically affects the voice signal. A previous work[1] with normal voices showed a high fidelity at high-bitrate compressions both in voice parameters and the amplitude-frequency spectrum. In the present work, dysphonic voices were tested through two studies. In the first study, spectrograms, long-term average spectra (LTAS), and fast Fourier transform (FFT) spectra of compressed and original samples of running speech were compared. In the second study, intensities, formant frequencies, formant bandwidths, and a multidimensional set of voice parameters were tested in a set of sustained phonations. Results showed that compression at high bitrates (96 and 128 kbps) preserved the relevant acoustic properties of the pathological voices. With compressions at lower bitrates, fidelity decreases, introducing some important alterations. Results from both works, Gonzalez and Cervera[1] and this paper, open up the possibility of using MPEG-compression at high bitrates to store or transmit high-quality speech recordings, without altering their acoustic properties.

Key Words: Pathological voice—Voice parameters—Speech compression—MPEG—mp3—MDVP—Acoustical analysis.

INTRODUCTION

As new technologies appear, it is important that their potential utility for the voice clinic and research be considered and that their validity be carefully assessed. In recent years, a revolutionary signal compression technique has had a great impact on the field of sound and music, and it would appear to be of use to the community of voice and speech specialists. 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[2] and Brandenburg[3]).

As occurs in music, high-quality recordings of speech signals require a considerable amount of data

Accepted for publication October 10, 2002.

From the *Universitat Jaume I, Castellon, Spain; †Universitat de Valencia, Spain.

Address correspondence and reprint requests to Julio Gonzalez, Department of Basic and Clinical Psychology and Psychobiology, Universitat Jaume I, Castellon, 12080 Castellon, Spain. E-mail: gonzalez@psb.uji.es

Journal of Voice, Vol. 17, No. 2, pp. 126–139

0892-1997/2003 $30.00+0

doi: 10.1016/S0892-1997(03)00007-9

storage. To attain the standard compact disk (CD) quality, the audio signal needs to be sampled 44,100 times per second and each sample requires a resolution of 16 bits; this gives 705 kbps, or 1410 kbps if stereo. For instance, one minute of a high-quality recording needs around 5 MB, or 10 MB if stereo (1byte equals 8 bits). This is a timely question because although the capacity of data storage has increased vastly in the last years, the teletransmission between distant laboratories is a true bottleneck. The collaboration among research centers around the world, frequently in cross-linguistic studies, is becoming ever more common. The use of the Internet imposes serious restrictions, and it will continue to do so for at least the next few years, in the interchange of high-quality speech signals. In this sense, the potentiality of a notable reduction of data size while preserving the signal quality should be considered. On the other hand, the possibility of drastically reducing the storage size of long high-quality voice files should not be discarded ahead of time.

Compression coding of digital data can be characterized as either "lossy" or "lossless." Systems of lossless audio compression (eg, *Shorten*, Cambridge University Engineering Department, Cambridge, UK, *DVD-Audio*, Dolby Laboratories Inc., San Francisco, CA, *MLP*, Meridian Audio Ltd., Stonehill, Huntingdon, UK, *WaveZip*, Gadget Labs Inc., Portland, OR, *Pegasus SPC*, Pegasus Imaging Corporation, Tampa, FL, *Sonarc*, Sonarc Audio Compression, Wilsonville, OR, *LPAC*, Communication System Group, Technische Universität Berlin, Germany among others) allow the exact reconstruction of the original signal, but they reach very low compression ratios, generally inferior to 2 : 1. MPEG-1 Layer 3 system, or mp3, is of the lossy variety; it is a sub-band coder that applies psychoacoustic coding schemes, removing parts of the signal that are perceptually irrelevant. Table 1 shows the ratios achieved by different degrees of compression of original voices digitized at the maximum values required for high-quality recordings, 50 kHz or 44.1 kHz. Each compression condition has an associated output sampling frequency. The different degrees of compression are indicated by the bitrate, or number of bits per second that will be contained in the encoded file, measured as kbps, or kilobits per second. One of the main goals in compression technology is the development of algorithms that preserve as much sound quality as possible even at very low bitrates. In general, the higher the bitrate, the higher the quality of the sound will be, but on the other hand, the file will be larger. MPEG-1 Layer 3 is an international ISO/MPEG standard—ISO/IEC 11172-3[4]— that achieves a very high-quality sound for middle and high bitrates. At these bitrates, trained listeners found it difficult to detect differences between original and compressed signals.[5,6] 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.

Listening tests[5,6] show that practically CD sound quality is obtained with MPEG-1 Layer 3 at 96 kbps per channel, whereas the size of the audio files is divided by 7-8. For more demanding musical pieces, such as piano concerts, it is advisable to increase the bitrate to 120 kbps. As regards the speech signal, results of these listening tests show that at 96 kbps or more, and in many cases even at 64 kbps, the compressed voice is audibly indistinguishable from the original.

This high perceptive efficiency is achieved by means of the application of a psychoacoustic model in the coding schema. MPEG-1 Layer 3 works by dividing the signal frequency spectrum into 32 subbands matching the psychoacoustic properties of the human ear in frequency resolution of the cochlea. For each sub-band, an algorithm calculates the perceptual masking effect caused by the other subbands. The masking effect raises the threshold of the noise floor, reducing the effective dynamic range of the signal. This reduced range requires fewer bits for codification, and this is the main opportunity 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 subband is lowered to $60 - 40 = 20$ dB, codified by just 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 to 5 ms) and after (up to 100 ms) a loud sound. This data reduction allows greater compression so

TABLE 1. *Compression Ratios of Original Speech Signals Digitized at 50 kHz or 44.1 kHz and Compressed at Different Bitrates*

	MPEG 128 kbps 44.1 kHz	MPEG 96 kbps 44.1 kHz	MPEG 64 kbps 44.1 kHz	MPEG 32 kbps 22050 Hz
Originals digitized at:				
50 kHz	6.3 : 1	8.3 : 1	12.5 : 1	25 : 1
44.1 kHz	5.5 : 1	7.4 : 1	11 : 1	22.1 : 1

Note: In each compression condition, bitrate (kbps, or kilobits per second) and output sampling rate are indicated.

as to be able to store or transmit audio signals without loss of sound quality.

Although MPEG appears to achieve the fidelity perceived by the human ear, we still lack precise information on the degree of distortion that MPEG compression could introduce in the voice signal. In our previous work[1] using normal voices—sustained phonations of /a/—we compared long-term average spectrum (LTAS) and a set of 29 multidimensional voice parameters between original and compressed voice signals at different bitrates. The results showed that at high-bitrate compressions, both voice parameters and amplitude-frequency spectra were very similar for uncompressed and compressed signals. Nevertheless, conclusions obtained with normal voices cannot be generalized to pathological voices. Compared to normal voices, audio signals corresponding to dysphonic voices are far more complex and are characterized by high variability of fundamental frequency and intensity, low signal-to-noise (SNR) ratio, voice breaks, and rapid shifts of vocal parameters that might be affected by the compression schema in ways that normal signals are not. Thus, we consider that the potential utility of MPEG compression in the voice clinic and research should be carefully assessed.

The aim of the present study is to ascertain the extent to which relevant acoustic properties of pathological voice signals are affected by MPEG-1 Layer 3 compression at different bitrates. In the present work, we extended our previous test of normal voices[1] to dysphonic voices, and a more comprehensive set of acoustical analyses was applied through two studies. In the first study, spectrograms, LTAS, and fast Fourier transform (FFT) spectra of compressed and original samples of pathological running speech were compared. In the second study, intensities, formant frequencies, formant bandwidths, and a multidimensional set of voice parameters of compressed and original samples were matched in a set of sustained phonations from dysphonic subjects.

STUDY I: RUNNING SPEECH

Method

Speaker

The speaker was a 33-year-old male bilingual speaker of Spanish and Catalan languages, who worked as an operator for a telephone company in Castellon (Spain). He was found to have dysphonia by hyperfunction, with bilateral vocal fold edema and frequent chorditis.

Apparatus

The recording was performed with a Shure SM58 microphone, at a distance of about 12 cm from the mouth, and a Sony-TCD D-8 digital audiotape (DAT) recorder with a sample frequency of 44.1 kHz. This frequency was chosen because this is the optimal output frequency for all the compression conditions applied (with the exception of the 32 kbps condition). The acoustical analysis was performed using Praat v.4.0,[7] a widely used[8,9] speech signal processing software application developed at the University of Amsterdam. This software was selected for its suitability and the quality of its graphic printing.

Voice sample

The subject read a passage of 115 words in Spanish taken from a novel by the Nobel Prize winner Camilo Jose Cela. This passage was selected because

it contains a broad variety of Spanish sounds. The passage was read at a normal speed and was recorded in a soundproof room. The recorded voice was transferred from DAT to PC with a Sound Blaster Platinum Live 5.0 sound card (Creative Technology Ltd., Singapore) over a RK-DA10 fiber optic cable and converted into a digital mono 16-bit WAV file sampled at 44.1 kHz.

Afterward, the voice signal was compressed by means of the Fraunhofer-Thomson compression scheme, which is the original and highest quality MPEG-1 Layer 3 algorithm, at the following bitrates and output sample frequencies: 128 kbps (44.1 kHz), 96 kbps (44.1 kHz), 64 kbps (44.1 kHz), and 32 kbps (22,050 Hz). These values give a wide set of compression rates ranging from 5.5 : 1 to 22.1 : 1 (Table 1). These output sample frequencies were those recommended by the authors of the algorithm for each bitrate. The compression scheme was implemented in the *Cool Edit 2000* program by Syntrillium Software Corporation (Phoenix, AZ).

Acoustic analysis

For each compression condition and original sample, we obtained the following analysis outputs comparison: (1) spectrogram of a sentence, (2) LTAS spectrum of the passage, and (3) FFT-spectrum of a vowel. Details of parameters used will be given in each section of results.

Results

Spectrograms

The first sentence of the passage, "Yo, señor, no soy malo" [literally: "I, sir, am not bad"], was selected and spectrograms were obtained from the original and from each of the compression conditions. The parameters selected were the following: Fourier method, frequency range display: 0-22 kHz, Gaussian window, analysis width: 5 ms, time step: 2 ms, frequency resolution: 260 Hz, pre-emphasis: 6 dB/octave. The Gaussian window was chosen, instead of the more common Hamming or Hanning windows, because according to the authors of the Praat program, it is superior, as it gives no sidelobes in the spectrograms.

Figure 1 shows the spectrograms of the original (A), and the compressed signal at 128 kbps (B), 96 kbps (C), 64 kbps (D), and 32 kbps (E). A visual inspection of them reveals two main effects of MPEG compression compared to the original signal: (1) A reduction of the effective signal frequency range: at 128 kbps, all the energy above 18.2-18.3 kHz is removed, and therefore not encoded; at 96 kbps, all the energy above 16.8-16.9 kHz is discarded; at 64 kbps, the same occurs above 11-12 kHz; and at 32 kbps, above 6.5-7.5 kHz. (2) In the effective frequency range, all perceptually irrelevant energy, according to the psychoacoustic algorithm applied, is also discarded. This effect, which is stronger at lower bitrates, acts mainly on the background noise of the signal, giving rise to the appearance of "gaps" where the original floor energy is very low. Therefore, both effects are two sources of a substantial saving in the amount of bits because no bit is used to encode the energy discarded.

Beyond these two effects, no other alteration is evident by visual inspection. For the purpose of studing quantitatively this question, two analyses (LTAS and FFT) were conducted to compare original and compressed signals. LTAS was applied to test if overal spectrum is conserved by compression through the complete frequency range of voice. FFT was used to study if the fine-grain structure of signal is preserved within the effective frequency range.

Long-term average spectrum

In order to study to what extent compression preserves the energy profile, dynamic range and overall tilt of the original voice, an LTAS was performed.

For each experimental condition, the LTAS was carried out through the complete passage to obtain a representation of the power spectrum expressed in decibels as a function of frequency. The same parameter values used to obtain the spectrograms were selected except that the frequency range display was 0-16,050 kHz. The analysis provided amplitude values at 150 Hz intervals, a total of 108 measurements being obtained for each condition, corresponding to the following frequencies: 0, 150, 300, 450, and so on up to 16,050 Hz.

Figure 2 shows the amplitudes in the 0-16,050 kHz range for each compression (and original) condition. To avoid negative values, all data were shifted up by 20 dB. We can observe that up to the point of 6800 Hz, the five lines in the figure are close together and parallel. At this frequency breakpoint,

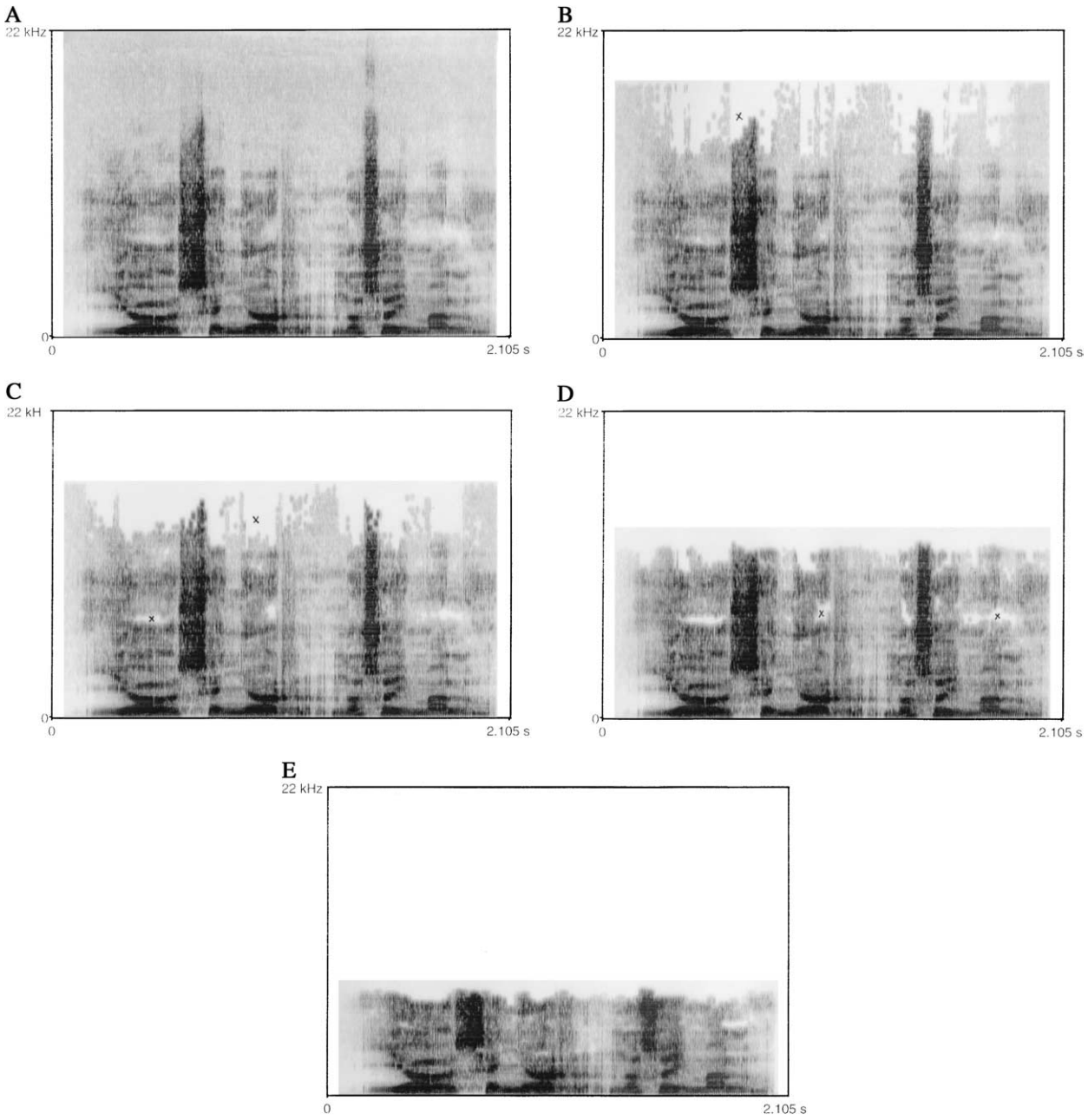

FIGURE 1. Spanish sentence: "Yo, señor, no soy malo" [literally: "I, sir, am not bad"]. Spectrograms of the original (A), and the compressed signal at 128 kbps (B), 96 kbps (C), 64 kbps (D), and 32 kbps (E). Some "gaps" where the energy is discarded are indicated (x).

the speech 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 11.5 kHz, we can see that the signal compressed to 64 kbps diverges strongly away from the others, as its energy level drops. At about 13 kHz, the signal compressed to 96 kbps begins to diverge away from the others,

although conserving a certain level of energy. Finally, the other two signals, original and compressed at 128 kbps, virtually maintain their parallelism throughout the whole of the 0-16,050 Hz frequency range. This parallelism is almost perfect because the mean difference with respect to the original (0.21 dB below) is maintained with barely any variation (range: −0.06 to +0.79 dB, excluding the two last intervals) in the whole of the frequency range. Thus, the Pearson correlation original versus MPEG-128 kbps across the 108 intervals is $r = 0.999$. To test the fitting between the original and the other compressed LTAS, Pearson coefficients were calculated across frequency intervals below each breakpoint, all conditions giving values equal or superior to 0.999.

In sum, all compression conditions preserved the long-term average spectrum to a high degree of accuracy, some a few tenths of a decibel under the original, along the following frequency ranges: 0-6800 Hz for 32 kbps, 0-11,500 Hz for 64 kbps, 0-13,000 Hz for 96 kbps, and 0-16,000 Hz for 128 kbps.

FFT spectrum

For each condition, an FFT was applied to get a spectrum of the Spanish vowel /o/ from the word "soy" [I am] belonging to the first sentence of the passage. Analysis was carried out on the first ten cycles of the vowel. Previously, a downsampling to 11,050 Hz was performed to restrict spectrum to the 0-5512.5-Hz range (Nyquist frequency). The parameters used were the default values recommended by the authors of Praat Software[7]: Fourier method, maximum frequency: 5 kHz, Gaussian window, analysis width: 5 ms, time step: 2 ms, frequency step: 20 Hz, pre-emphasis: 6 dB/octave.

In each FFT, a total of 513 values expressed in decibels as a function of frequency were obtained. Figure 3 shows the graphic representation of all spectra. Visually, the high fidelity to original FFT spectrum across conditions is obvious, especially in the higher bitrates (MPEG-128 and MPEG-96). The similarity of spectra with the original is given as much in the harmonic structure as in the noise of the high-frequency region. The absolute mean discrepancies (AMD) and Pearson correlations original versus compressed signals across the 513 values were the following: MPEG-128 (AMD = 0.56 dB, SD = 1.22, $r = 0.994$); MPEG-96 (AMD = 0.88 dB, SD = 1.36, $r = 0.991$); MPEG-64 (AMD = 2.23 dB, SD = 2.55, $r = 0.962$); MPEG-32 (AMD = 2.79 dB, SD = 3.17, $r = 0.942$). Data show a very good fit in the two higher bitrates, and less good in the two lower bitrates.

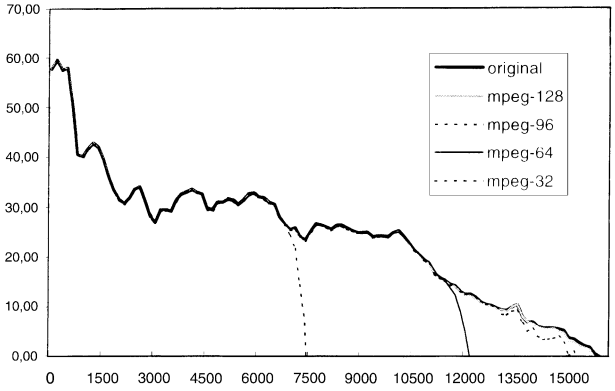

FIGURE 2. LTAS of original and compressed voice signals from running speech. Representation of the mean values of amplitude (in decibels) in each frequency level analyzed in the range 0-16,050 Hz.

STUDY II: SUSTAINED PHONATION

In the second study, compressions were tested by means of relevant analysis in the voice clinic and of research across several dysphonic speakers. Besides correlations, AMD between original and compressed values was chosen as the statistic for comparative purposes. AMD is defined as the mean of the original-compressed differences taken in absolute values. Two very different case-to-case distributions could yield close overall means because positive differences may be canceled out by the negative ones. The use of AMD prevents this possibility because all differences are accumulated, which is why it is preferable in studies where parameter values are compared through several subjects (see, eg, Kent and Duffy[10]).

Method

Voice samples

Voice samples were selected from the Voice Disorder Database of Kay, recorded at the Massachusetts Eye and Ear Infirmary.[11] Selected samples were

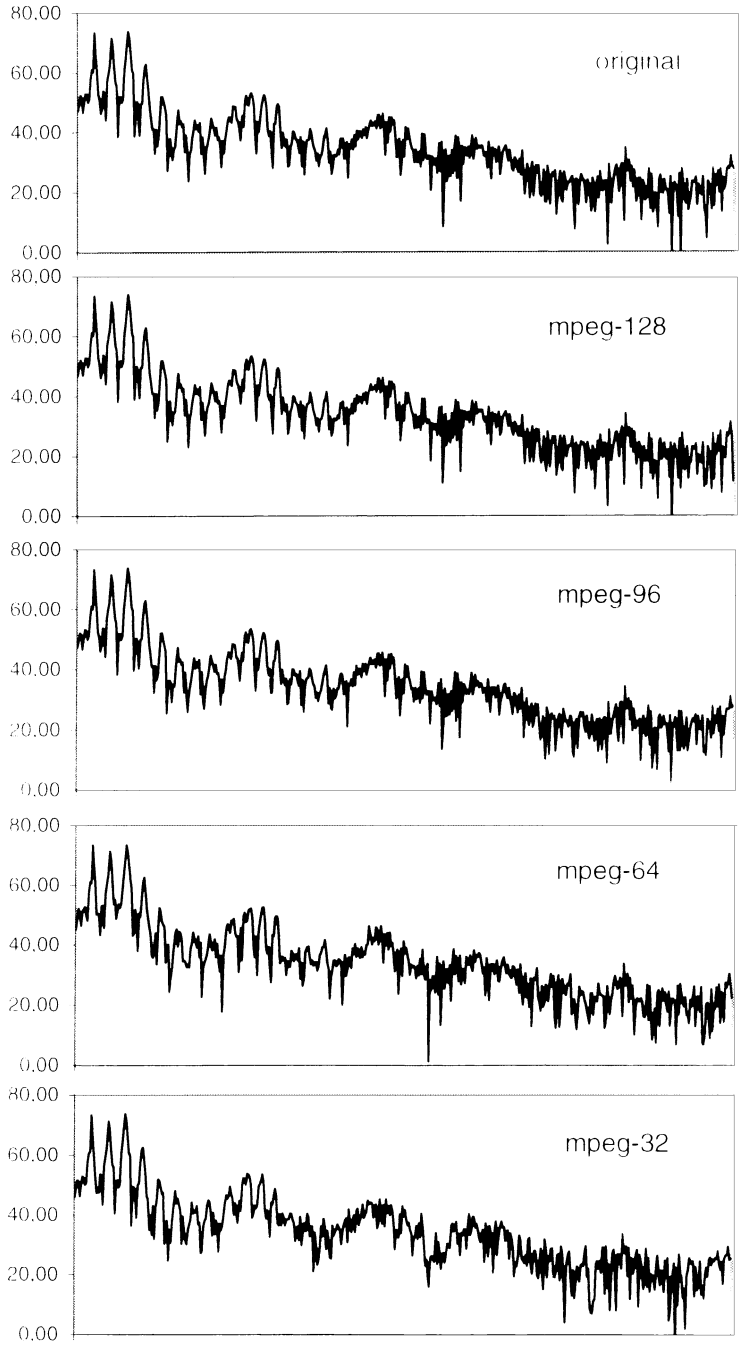

FIGURE 3. FFT of original and compressed samples of the Spanish vowel /o/.

sustained phonations of /a/ vowel lasting approximately 1 second provided by 17 pathological speakers, 9 males and 8 females, who had a variety of voice disorders (see the Appendix). All voices had an apparent periodic structure. Ten patients fell within the type 1 of the Denver classification,[12] and seven (Crm12, Jpp27, Mcb20, Mrc20, Pat10, Pmc26, Wxe04) within the type 2. All voice samples were recorded on DAT tape in a soundproof room at the Massachusetts Eye and Ear Infirmary at a sampling

rate of 44.1 kHz. From the DAT tape, the recordings were converted into an analog signal and acquired into a CSL-Kay system model 4300B at sampling rates of 50 kHz (with 24 kHz anti-aliasing filtering). They were then saved as CSL files (.NSP format).

The 17 NSP files were converted to WAV format and compressed following the same procedure as in Study I. Bitrates and output sample frequencies were 128 kbps (44.1 kHz), 96 kbps (44.1 kHz), 64 kbps (44.1 kHz), and 32 kbps (22050 Hz). From originals digitized at 50 kHz, these bitrates gave a wide set of compression rates ranging from 6.3:1 to 25.1:1 (Table 1). All output sample frequencies were the optimal recommended by the algorithm authors for each bitrate. Furthermore, given that all compression options also involve downsampling, an additional condition was stipulated for comparative purposes: ie, that sample frequency of the original signal is converted from 50 to 44.1 kHz without there being any MPEG compression. This condition will be called the "downsampled" condition.

Acoustic analysis

For each original /a/ vowel and each compressed (downsampled) version the following analysis were performed: (1) Calculation in decibels of total signal intensity. (2) LPC analysis to get the frequency and bandwidth of the first four formants. (3) Extraction of a set of ten multidimensional voice parameters. The two first analyses were carried out with the Praat software. Details of settings used will be given in each section of results.

Results

Intensity

The mean intensity value, expressed in decibels, was obtained for each signal. All the correlations between intensities of original versus compressed signals were perfect (see Table 2). The intensities of downsampled versions yielded 0.01 dB as AMD of the original. The AMD original-compressed signals were only about 0.14B0.21 dB, but all them were significantly greater than AMD original-downsampled. All standard deviations across speakers are very low, which is consistent with the perfect correlations. On the other hand, it is necessary to highlight the difference between both higher bitrates (AMD: 0.14-0.15 dB) and both lower ones (AMD: 0.21 dB).

Formants

Frequency (F1-F4) and bandwidths (B1-B4) of the first four formants of each vowel signal were calculated. Prior to analysis, all voice files were downsampled to 11,050 Hz. Analysis was based on the LPC (Burg) algorithm of Press et al,[13] with the following parameters: number of poles, 10; maximum number of formants, 5; maximum formant, 5500 Hz.; Gaussian-like window; analysis width, 25 ms; pre-emphasis from 50 Hz.

As a first approach, in each compression condition (including the only downsampled signal condition), two multivariate analysis of variance (MANOVA) were performed considering separately F1-F4 and B1-B4 values as dependent variables and original versus compressed signal as factor. MANOVA (see Table 3) only showed a marginal effect in formant frequencies due to compression of voice signal at a bitrate of 32 kbps [$F(1,16) = 3.09$, $p = 0.098$].

Mean values of formant frequencies/bandwidths of originals and AMD original-compressed (downsampled) signals across speakers are shown in Table 2. Pearson correlations with originals are very high for the downsampled condition and compressions at two higher bitrates. All coefficients were higher than 0.96; the only exception was B1 for the downsampled condition, because of a discrepancy of 1014 Hz in a speaker (Mcb20). As in the "downsampled" condition, the signal was not compressed, this condition was chosen as control to compare the compression effects. Consequently, AMD original-downsampled signals were compared to AMD original-compressed signals. All comparisons were made of one tail, because if any difference exists, this would be in the sense of more discrepancy in compressed conditions. Data show that no discrepancy for compressions at two higher bitrates was significantly greater than for the downsampled condition, except for F4 at MPEG-96 kbps. On the contrary, most of the discrepancies observed in MPEG-64 and MPEG-32 kbps were bigger than in the downsampled condition, and some correlations were below 0.90.

Voice parameters

A set of ten voice parameters were compared across original versus compressed (downsampled)

TABLE 2. *Comparison of Intensities, Formant Frequencies (F1-F4), and Bandwidths (B1-B4) between Original and Compressed Speech at Different Bitrates Through 17 Dysphonic Subjects*

Mean (sd) values from original pathological voices	**Downsampled**	**MPEG 128 kbps**	**MPEG 96 kbps**	**MPEG 64 kbps**	**MPEG 32 kbps**
	AMD (sd) and Pearson correlations				
Intensity (dB) 76.56 (2.94)	**0.01 (0.01) $r = 1$**	**0.14 (0.05)*** $r = 1$**	**0.15 (0.02)*** $r = 1$**	**0.21 (0.03)*** $r = 1$**	**0.21 (0.04)*** $r = 1$**
Formants (Hz):					
F1 = 675 (116)	3.53 (4.43) $r = 0.999$	4.00 (7.95) $r = 0.997$	8.35 (14.88) $r = 0.991$	7.65 (10.37)* $r = 0.996$	42.59 (89.63)* $r = 0.704$
F2 = 1354 (258)	5.29 (9.14) $r = 0.999$	9.18 (13.82) $r = 0.998$	9.94 (18.17) $r = 0.997$	14.35 (18.90)* $r = 0.997$	35.18 (74.21) $r = 0.954$
F3 = 2762 (225)	4.41 (5.05) $r = 1$	7.59 (10.68) $r = 0.998$	14.18 (18.34) $r = 0.995$	12.47 (13.23)** $r = 0.997$	48.06 (106.19) $r = 0.886$
F4 = 3801 (239)	11.53 (20.55) $r = 1$	9.82 (12.41) $r = 0.998$	17.18 (14.97)* $r = 0.997$	27.00 (38.57)** $r = 0.985$	47.29 (64.26)* $r = 0.953$
B1 = 424 (288)	64.35 (244.83) $r = 0.791$	18.71 (24.91) $r = 0.997$	15.18 (32.97) $r = 0.993$	89.82 (144.85) $r = 0.874$	51.24 (84.29) $r = 0.939$
B2 = 248 (138)	11.65 (27.18) $r = 0.986$	12.59 (27.51) $r = 0.979$	16.76 (27.38) $r = 0.976$	29.94 (48.43)** $r = 0.950$	26.35 (22.48)* $r = 0.968$
B3 = 423 (440)	15.41 (25.04) $r = 0.998$	30.00 (62.94) $r = 0.988$	30.53 (54.43) $r = 0.990$	35.06 (38.48)* $r = 0.995$	46.76 (84.85) $r = 0.980$
B4 = 667 (804)	75.59 (153.39) $r = 0.990$	98.76 (291.03) $r = 0.969$	88.06 (262.27) $r = 0.984$	224(401.44)* $r = 0.836$	222.82 (355.88)* $r = 0.861$

Note: Shown for each condition are absolute mean discrepancy (AMD) (standard deviations between parenthesis) and Pearson correlations (r) between values from original and compressed signals. Downsampled condition is included.

*Mean difference between AMD original-downsampling vs AMD original-MPEG condition at a significance value $p < .05$ (t test, one tail).

**Idem at a significance value $p < .01$ (t test, one tail).

***Idem at a significance value $p < .001$ (t test, one tail).

versions. Parameters were from the *Multi-Dimensional Voice Program* (MDVP) of Kay Elemetrics Corp. We chose this program because it is an important analytical tool increasingly used in voice studies.[10,14,15] The ten parameters were selected from the total MDVP set to deal with relevant voice dimensions, avoid some redundancies, eg, absolute and relative jitter, and discard others, tremor parameters, with very low test-retest reliability.[16] Briefly, the parameters were the following. Fundamental Frequency parameter: *Average Fundamental Frequency* (*Fo*). Frequency perturbation parameters: *Jitter Percent* (*Jitt*) /%/: relative period-to-period variability of the pitch period; *Relative Average Perturbation* (*RAP*) /%/: introduced by Koike,[17] this parameter gives the relative evaluation of the period-to-period variability of the pitch with a smoothing factor of three periods. Amplitude perturbation parameters: *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,[18] which gives the relative evaluation of the variability of the peak-to-peak amplitude at smoothing of 11 periods; this smoothing reduces the sensitivity of APQ to pitch extraction errors. Noise parameters: *Noise to Harmonic Ratio* (*NHR*): a general evaluation of the noise presence in the analyzed signal; this is the ratio of inharmonic energy in the range 1500B4500 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 in a 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. This parameter is not actually a measurement of noise, but its formula is similar to the above two parameters and is listed in the same category in the MDVP manual. Parameters of Subharmonic components: *Number of Subharmonic Segments* (*NSH*). Parameters of Voice irregularities: *Number of Unvoiced Segments* (*NUV*).

Reliability of the measures. Our values obtained from the original samples were compared to the values provided by the Disordered Voice Database,[11] and Pearson correlations were calculated. All subjects obtained coefficients higher than 0.99 across the ten selected parameters. All parameters yielded coefficients higher than 0.91 across the subjects. Average perturbation measures were quite high because of three subjects with extreme values (Crm12, Jpp27, and Pat10). For example, mean Jitt value (2.70%) is far apart from the threshold for normal voices supplied by the MDVP Manual (1.04%), but it decreases to 1.67% if the three extreme subjects are excluded. Likewise, means of RAP (1.57%), Shim (7.01%), and APQ (5.09%) drop to 0.98, 5.97, and 4.48%, respectively.

Previous research[10] has revealed that some voice parameters are very sensitive to small variations of the input signal. These authors found that analysis on repeated edits of the same digitized signal may yield different values, so that a certain variability is expected in our study, even in the downsampled condition. As a first approach, in each compression (downsampled) condition, a MANOVA was performed considering the voice parameter values as dependent variables and original versus compressed signal as factor. MANOVA (see Table 3) only showed a significant effect due to compression of voice signal at 32 kbps bitrate [$F(1,16) = 6.46$, $p = 0.022$].

Table 4 shows the mean values through speakers of original parameters and AMD original-compressed (downsampled) signals. Also, Pearson product moments were calculated between original and compressed (downsampled) values. Data show correlations higher than 0.90 for downsampled and MPEG-128 and MPEG-96 in all parameters except for RAP and APQ. In general, no clear difference in AMD or correlations between downsampled versus two higher bitrates is apparent. Some correlations are even greater in MPEG-128 or MPEG-96 than in downsampled condition (Jitt, RAP, Shim, APQ). When original-MPEG-128 or original-MPEG-96 discrepancies are statistically larger than original-downsampled discrepancies (VTI, SPI, NUV), the differences, in general, are small with regard to the actual values of original. On the other hand, parameters in MPEG-64 and, mainly, MPEG-32

Journal of Voice, Vol. 17, No. 2, 2003

TABLE 3. *Results of MANOVA in Each Condition with Formant Frequencies (F1-F4), Formant Bandwidths (B1-B4), or Voice Parameters as Dependent Variables and Original vs. Compressed (Downsampled) Speech Signal as Factor*

Dependent Variables	Downsampled	MPEG 128 kbps	MPEG 96 kbps	MPEG 64 kbps	MPEG 32 kbps
F1 F2 F3 F4	$F(1,16) = 0.83$ $p = 0.377$	$F(1,16) = 0.06$ $p = 0.814$	$F(1,16) = 0.01$ $p = 0.916$	$F(1,16) = 2.70$ $p = 0.120$	$F(1,16) = 3.09$ $p = 0.098$
B1 B2 B3 B4	$F(1,16) = 1.77$ $p = 0.201$	$F(1,16) = 1.23$ $p = 0.285$	$F(1,16) = 0.66$ $p = 0.427$	$F(1,16) = 1.49$ $p = 0.240$	$F(1,16) = 0.07$ $p = 0.792$
Voice parameters	$F(1,16) = 1.74$ $p = 0.206$	$F(1,16) = 0.43$ $p = 0.519$	$F(1,16) = 0.01$ $p = 0.980$	$F(1,16) = 0.28$ $p = 0.601$	$F(1,16) = 6.46$ $p = 0.022$

TABLE 4. *Comparison of Voice Parameters between Original and Compressed Pathological Voices at Different Bitrates*

Mean (sd) values from original pathological voices[a]	Downsampled	MPEG 128 kbps	MPEG 96 kbps	MPEG 64 kbps	MPEG 32 kbps
	AMD (sd) and Pearson correlations				
Fo (Hz) = 142.50 (42.69)	1.01 (3.59) $r = 0.997$	0.37 (0.84) $r = 1$	1.34 (3.64) $r = 0.997$	4.41 (15.19) $r = 0.935$	0.51 (1.23) $r = 1$
Jin (%) = 2.70 (2.82)	0.49 (1.30) $r = 0.903$	0.53 (1.03) $r = 0.912$	0.50 (0.72) $r = 0.956$	0.32 (0.44) $r = 0.988$	0.64 (1.37) $r = 0.859$
RAP (%) = 1.57 (1.60)	0.30 (0.77) $r = 0.885$	0.35 (0.65) $r = 0.890$	0.29 (0.41) $r = 0.953$	0.21 (0.28) $r = 0.987$	0.36 (0.78) $r = 0.857$
Shim (%) = 7.01 (3.28)	0.59 (1.82) $r = 0.903$	0.36 (0.68) $r = 0.977$	0.86 (1.34) $r = 0.943$	1.06 (1.68)*** $r = 0.905$	0.93 (1.10) $r = 0.944$
APQ (%) = 5.09 (2.14)	0.45 (1.64) $r = 0.876$	0.23 (0.34) $r = 0.986$	0.61 (1.14) $r = 0.935$	0.78 (0.92) $r = 0.937$	0.93 (0.92) $r = 0.960$
NHR = 0.211 (0.107)	0.006 (0.020) $r = 0.990$	0.004 (0.008) $r = 0.998$	0.022 (0.046) $r = 0.956$	0.014 (0.021)*** $r = 0.992$	0.019 (0.028)* $r = 0.970$
VTI = 0.102 (0.061)	0.005 (0.008) $r = 0.994$	0.010 (0.012) $r = 0.969$	0.011 (0.013)* $r = 0.963$	0.017 (0.023)* $r = 0.905$	0.032 (0.035)** $r = 0.894$
SPI = 9.42 (4.71)	0.13 (0.25) $r = 0.998$	0.35 (0.33)* $r = 0.997$	0.44 (0.47)* $r = 0.993$	0.56 (0.65)* $r = 0.986$	1.58 (2.23)** $r = 0.887$
NSH = 1.71 (3.37)	0.18 (0.53) $r = 0.986$	0.35 (0.70) $r = 0.976$	0.29 (0.59) $r = 0.982$	0.71 (1.69) $r = 0.868$	0.47 (1.37) $r = 0.923$
NUV = 1.82 (3.30)	0.18 (0.39) $r = 0.992$	0.47 (0.80)* $r = 0.973$	0.41 (0.93) $r = 0.952$	0.47 (0.62)* $r = 0.977$	0.71 (0.92)** $r = 0.970$

[a]Parameter abbreviations are explained in the text.

Note: Downsampled condition is included. See footnote for Table 2.

compressions show more important alterations in comparison with only downsampled signals.

GENERAL DISCUSSION

In a previous study[1] it was found that normal speech compressed at high bitrates by means of MPEG Layer 3 schema preserved relevant acoustic properties. Results showed a high fidelity both in voice parameters and the amplitude-frequency spectrum of compressed voices. These observations not could be directly extrapolated to pathological speech because dysphonic signals might be affected by the compression schema in ways that normal signals are not. On the other hand, it was interesting to extend the scope of the test to a wide set of acoustical correlates commonly used in the research and clinical practice.

The results of the two studies presented here show that for high bitrates, such as 128 or 96 kbps, compressed pathological voice also shows high fidelity to original signal. Starting from high-quality recordings of speech at 50 kHz, or 44.1 kHz, it is possible to reach compression rates of about 6-8 : 1 without losing hardly any signal quality. Listening tests[5,6] had demonstrated that music and speech compressed at high bitrates was virtually indistinguishable from original. In these two works (present and Gonzalez and Cervera[1]) objective analyses show that the main effect of MPEG-compression in normal and pathological speech is on the nonaudible ground noise, hardly affecting the quality of signal. High-bitrate compression totally discards any energy from a frequency band perceptually irrelevant—above 18 kHz for MPEG-128, and above 16.8 kHz for MPEG-96 kbps—reducing the effective frequency broadband to a useful range. Moreover, below this frequency limit, compression also removes very weak and irrelevant ground energy. Both effects are the main sources of bit saving in the process of signal encoding. Nevertheless, spectrograms reveal that perceptually relevant noise energy such as consonant bursts, aspirations, frictions, and weak breaths between utterances appears to be preserved by the compression schema, at least at high bitrates.

Spectrograms and FFT-spectra show that, for high-bitrate compressions, both harmonic and nonharmonic structure of speech is well preserved below a frequency breakpoint. This breakpoint where compressed energy profile begins to diverge from the original, as the long-term average spectrum shows, is sufficiently high to preserve all bandwidth of interest in speech signal (about 0-16,000 Hz for 128 kbps, and 0-13,000 Hz for 96 kbps). This structure preservation is corroborated by results from formant frequencies and bandwidths, calculated by means of an LPC algorithm: High-bitrate compressions do not really alter their values more than the mere downsampling of signal does.

One way to test the preservation of the fine-grain structure of compressed pathological speech is by means of a wide set of voice parameters. Automatic voice analysis with the MDVP of Kay has demonstrated its utility and reliability with pathological voices in other studies.[10] Nevertheless, voice parameters, because of the nature of the algorithms involved, frequently show important changes from minimal variations in the input signal, especially in pathological voices. For example, Kent and Duffy,[10] working with dysarthric voices, found important variations in some parameters measured from two editings of the same sustained vowel. For this reason, it is not uncommon for any manipulation of signal to give rise to some variations in the parameter values, even from a mere downsampling. Now then, our data indicate roughly that voice parameters are not much more sensitive to high-bitrate compressions than to a signal downsampling from 50 to 44.1 kHz. Compared with the previous study on normal voices,[1] compressed parameters differ from original more when they are pathological voices, but this also occurs in the only downsampled condition, and no selective major change due to compression is seen. Gonzalez and Cervera[1] found that the compression schema introduced a very tiny systematic 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 is only downsampled and are much lower than in the case when directly digitized samples are compared to taped voice samples; Gelfer and Fendel[19] found a variation of approximately 3.2 Hz and a correlation $r = 0.989$. The classic frequency perturbation parameters used in research and the speech clinic, such

as jitter, whether it is measured directly or by the widely used relative average perturbation (RAP) with a smoothing factor of three periods, are not more affected by compression at 128 or 96 kbps than by merely downsampling. The same is true regarding amplitude perturbation measures (Shim and APQ), especially in the case of MPEG-128. Contrary to what happens with jitter, Gelfer and Fendel[19] 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 ($r = 0.481$). Data from Gonzalez and Cervera[1] showed correlations of 0.999 in all the parameters comprising this class. Our current data with dysphonic signals yield correlation coefficients above 0.930 in MPEG-128 and MPEG-96 conditions. Even noise parameters, such as Noise to Harmonic Ratio (NHR), generally quite sensitive to any signal manipulation that could increase the noise level, does not seem to be especially affected by MPEG compression. In a recent study on the suitability of minidisk (MD) recordings for voice perturbation analysis, Winholtz and Titze[20] 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 SNR, which was approximately 10 dB less for MD recordings than for normal DAT recordings. In accordance with our values, the MPEG compression at high bitrates gives a much better SNR relationship than do MD recordings.

Everything stated above is true mainly for compressions at higher bitrates: 128 and 96 kbps. When compression is performed at lower bitrates, fidelity decreases, introducing more important alterations in the voice signal. Working at low bitrates, if the encoder runs out of bits, it will not encode some blocks of signal data with the required fidelity, which will have consequences in the fine-grain structure of the sound wave. In summary, according to Gonzalez and Cervera's[1] and current data, MPEG-compression at high bitrates does not seem to introduce major alterations that may degrade or deteriorate normal or pathological speech signal much more than a mere downsampling does. All tests carried out in both papers open up the possibility of using MPEG-compression at high bitrates to store or transmit high-quality speech recordings, without altering relevant acoustic properties. Nevertheless, this statement could not be extended, without further specific research, to pathological voices with very degraded harmonic structure, such as oesophageal speech,[21,22] or signal with no apparent periodic structure, type 3 of Denver classification.[12]

Acknowledgments: This work was supported by *Fundació Caixa Castelló-Bancaixa* and the Universitat Jaume I, Castellon, Spain, Project P1.1A2002-01.

REFERENCES

1. Gonzalez J, Cervera T. The effect of MPEG audio compression on a multi-dimensional set of voice parameters. *Log Phon Vocol*. 2001;26:124–138.
2. 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;42:780–792.
3. Brandenbug K. MP3 and AAC explained. In: *Audio Engineering Society 17th International Conference on High Quality Audio Coding*; Florence, Italy, 1999.
4. 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, 1993.
5. ITU-R. Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. ITU-R Recommendation BS.1116, Geneva, Switzerland, 1994.
6. EBU. Basic audio quality requirements for digital audio bit-rate reduction systems for broadcast emission and primary distribution. CCIR document number TG 10-2/3, 1991:28.
7. Boersma P, Weenink D. Praat: A system for doing phonetics by computer, 2000. Available at: http://www.fon.hum.uva.nl/praat/.
8. Sebastian-Galles N, Dupoux E, Costa A, Mehler J. Adaptation to time-compressed speech: Phonological determinants. *Percept Psychophys*. 2000;62:834–842.
9. Fitch T, Reby D. The descended larynx is not uniquely hyman. *Proc R Soc London B*. 2001;268:1669–1675.
10. 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;8:129–136.
11. Massachusetts Eye and Ear Infirmary. *Voice Disorders Database, version 1.03* [CD-ROM]. Lincoln Park, NJ: Kay Elemetrics, 1993.
12. Titze IR. Workshop on Acoustic Voice Analysis. Summary Statement. Denver, CO: National Center for Voice and Speech; 1994.

13. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. Numerical Recipes in C: The art of scientific computing. Cambridge, UK: Cambridge University Press; 1992.
14. Cimino AM, Sapienza C. Reliability of the Multidimensional Voice Program (MDVP) for Acoustic Analysis of the Normal Voice. Paper presented at the 28th Annual Symposium of the Voice Foundation on the Care of the Professional Voice, Philadelphia, June 1999.
15. Corina J, Hilgers FJM, Verdonck-de-Leeuw IM, Koopmans-van Beinum FJ. Acoustical analysis and perceptual evaluation of tracheoesophageal prosthetic voice. *J Voice*. 1998;12:239–248.
16. Gonzalez J, Cervera T, Miralles JL. Anàlisis acùstico de la voz: Fiabilidad de un conjunto de paràmetros multidimensionales. [Acoustic voice analysis: Reliability of a set of multi-dimesional voice parameters]. *Acta Otorrinolaringol Esp*. 2002;53:256–268.
17. Koike Y. Application of some acoustic measures for the evaluation of laryngeal dysfunction. *Stud Phonol*. 1973; 7:17–23.
18. Koike Y, Takahashi H, Calcatera T. Acoustic measures for detecting laryngeal pathology. *Acta Otolaryngol*. 1977; 84:105–117.
19. Gelfer MP, Fendel DM. Comparisons of jitter, shimmer, and signal-to-noise ratio from directly digitized versus taped voice samples. *J Voice*. 1995;9:378–382.
20. Winholtz WS, Titze IR. Suitability of minidisc (MD) recordings for voice perturbation analysis. *J Voice*. 1998; 12:138–142.
21. Cervera T, Miralles JL, Gonzalez J. Acoustical Analysis of Spanish vowels produced by larigectomized subjects. *J Speech Lang Hearing Res*. 2001;44:988–996.
22. Miralles JL, Cervera T. Voice intelligibility in patients who have undergone laryngectomies. *J Speech Hearing Res*. 1995;38:564–571.

APPENDIX. *Description of the Pathological Voice Samples Used for the Study II*

File Name	Age	Gender	Diagnosis	Location of disorder
AJM05AN.NSP	70	M	hyperfunction, paralysis	unilateral left
CAK25AN.NSP	47	F	hyperfunction, ventricular vocal folds (mild), vocal fold edema	unilateral left
CRM12AN.NSP	50	M	Parkinson's disease, choreaic movements, hyperfunction, A-P squeezing (severe), ventricular compression (severe)	
DMP04AN.NSP	31	F	hyperfunction, A-P squeezing (mild), ventricular compression (moderate), polypoid degeneration (Reinke)	bilateral
EX104AN.NSP	33	M	gastric reflux, hyperfunction, immediate post surgery, interarytenoid hyperplasia. polypoid degeneration (Reinke). A-P squeezing (severe), ventricular compression (Severe), vocal fold polyp	bilateral
HX129AN.NSP	54	M	hyperfunction, A-P squeezing, interarytenoid hyperplasia	
JPP27AN.NSP	42	F	Blunt trauma, hyperfunction, laryngeal trauma, scarring hematoma, paralysis	bilateral unilateral left
JTM05AN.NSP	47	M	hemmoragic Reinke's edema	unilateral left
LAR05AN.NSP	61	M	hyperfunction, laryngocele, A-P squeezing (severe), ventricular compression (severe), mass	unilateral right
MCB20AN.NSP	74	F	hyperfunction, lymphode hyperplasia, A-P squeezing (severe), ventricular compression (severe), vocal fold polyp	unilateral right
MRC20AN.NSP	40	F	hyperfunction, A-P squeezing, vocal nodules	bilateral
NJS06AN.NSP	21	F	hyperfunction, A-P squeezing (mild), vocal nodules	bilateral
PAT10AN.NSP	33	M	anterior mass, hyeprfunction, A-P squeezing (moderate), post irradation, post vocal fold stripping, ventricular compression (severe) vocal fold atrophic vocal fold edema	 unilateral right unilateral left
PMC26AN.NSP	37	M	varix	unilateral right
RXS13AN.NSP	78	F	gatrix reflux, A-P squeezing (mild), ventricular compression (mild), white debris/patches	bilateral
SHC07AN.NSP	20	F	hyperfunction, A-P squeezing, vocal nodules	bilateral
WXE04AN.NSP	36	M	atrophic laryngitis, hyperfunction, mass, A-P squeezing (severe), ventricular compression	