Proceedings of the 5th WSEAS Int. Conf. on Signal Processing, Robotics and Automation, Madrid, Spain, February 15-17, 2006 (pp195-200)

Quality Control in the Radio Nacional de España Sound Archive Digitization Process

ROBERTO GIL-PITA[1], RAÚL VICEN-BUENO[1], ENRIQUE ALEXANDRE-CORTIZO[1]
JUAN C. GÓMEZ-CORNEJO[2], MIGUEL J. RODEÑO ARRÁEZ[3]
[1]Dpto. de Teoría de la Señal y Comunicaciones - Universidad de Alcalá – Madrid (SPAIN)
[2]IBM Global Services (SPAIN)
[3]Dpto. de Ciencias de la Computación - Universidad de Alcalá – Madrid (SPAIN)

Abstract: - The 20th century Spanish sound history can now be consulted online. This is a pioneer project in the broadcasting industry around the world. The Radio Nacional de España (RNE) sound archive has been massively digitized and several applications to access this information online have been developed. This archive is considered the most important audio archive in Spanish language in the world. Special innovation has been used in developing a statistical model to estimate the probability of error in the digitisation process, for a given confidence probability. Several kinds of errors have been considered, like human errors, mechanical errors, software errors or support errors. The results of the statistical model allows to say that the probability of error is lower than 0.3% for a confidence probability of 99.999%.

Key-Words: - radio digitization, quality control, multimedia context protection, online content.

1 Importance of the audio digitization projects and its quality control

UNESCO [1] "urges its Member States to assume responsibility of ensuring the identification, preservation and transmission to future generations of the digital heritage produced on their territory, and to promote the use of open standards in conjunction with the IT industry"

Long-term preservation of audiovisual data is hopeless: the carriers are unstable, the commercial lifetimes of the formats seem to become shorter and shorter, and the amount of data to be stored increases every day. The solution lies with automatically accessible, self controlling and self re-generating archival systems, also known as digital mass storage systems (DMSS). The features of such systems are [2, 3, 4]:

- The management of audiovisual data as computer files in mass storage systems, e.g. libraries or robotics of magnetic tape cartridges.
- An open file architecture to accommodate all audiovisual data together with catalogue/content information and written text (metadata).
- The access time of such systems is not of major importance.
- Data integrity is controlled automatically, and copying of the information onto new carriers is done automatically before mistakes cannot be fully corrected.
- Once new storage media and systems are available due to technical development, automated migration will be implemented.

Similar observations are done in IASA [5], a ten sheet must-read paper for this topic. The benefits of using a networked radio station with digital mass storage versus using a conventional radio sound archive are broadly described in [4, 6].

During the digitisation process several kind of errors may appear. The identification of sound archives may be erroneous, the source material may be damaged, and some errors may appear in the mechanical selection of the source supports. Furthermore, digital formats have a limited resolution set by the defined and finite sampling frequency and digital word length.

Digital resolution, as well as digital representation, must be controlled in order to guarantee a high quality of the digital sound material.

In principle, high digital resolutions are desirable for the adequate representation of all the minute subtleties of original analogue signals. It is important to understand that the intended signal is only part of a given sound document. The unintended and undesirable artefacts (noise, clicks, distortions) are also part of the sound document, even if they have been subsequently added to the original signal by mishandling or poor storage. Both have to be preserved with utmost accuracy.

When the amount of sound material is enormous, like in radio sound archives, it is impossible to check all the digitised files in order to guarantee that no errors are produced. So, some kind of quality control should be introduced in the digitisation process. This quality control must provide the possibility of detecting all the kinds of considered errors.

Unfortunately, even when some quality control is included during the digitisation process, it is impossible to check all the digitised files. So, only a few subset of digitised files can be tested by the quality control. The quality control can not guarantee that all the sound files are correctly digitised, but can provide useful information about the digitisation process in order to estimate the number of errors after the process.

Two quality control stages have been included in the Radio Nacional de España sound archive digitisation process. The information provided by these stages has been used as input to a statistical model that allows to estimate the error probabilities of the digitisation process.

The appearance of errors in a sound archive digitization has already been analysed in [6], where errors are classified as errors caused by recording and rerecording equipment, errors caused by medium (for example, tape or record), and error caused by operator. A system for real-time audio analysis of the digital data stream is incorporated to the digitization process, that generates reports and statistical measurements. The sound archive digitization project described in [6] does not include a model of the detection of errors that would permit to estimate the probability of erroneous digitization. This paper tackles this problem.

A survey done with ten broadcasters in Europe [7] indicates that the audio holdings are mainly in quarter inch tapes and shellac & vinyl discs, besides cassettes, DAT, CD, minidisks and Tandberg QIC cartridges.

2 RNE sound archive

The RNE sound archive has a very high historical value that includes 1,400,000 sound records that are equivalent to 190,000 recorded hours and that were located in thousands of square meters in linear shelves in the Radio House cellar in Madrid.

This archive is part of the Spanish and European national heritage because it has oral testimonies, as well as broad recorded performances and concerts for the 20th century historical, cultural and society study. This archive is considered the most important audio archive in Spanish language in the world.

The RNE digitized archive contents include items previous to the radio broadcasting in Spain, such as the testimonies of the Austrian Emperor Francis Joseph (1889), Thomas Edison (1898), Emperor William II, first world war declaration and Lenin speaking to the attendees at the III International (1919). In addition, the contents include many oral testimonies of key persons in 20th century. So it may be noticed the different Hitler voice tone in 1936 and in his last speech in 1945. Also it includes many classical music concerts recorded live by RNE, as well as a broad collection of international pop music.

3 Project objectives and description

The three objectives of the projects are:

- Preserve the entire historical heritage by digitizing the contents. Although the RNE's security and environmental control systems guarantee the optimal conditions in the preservation of this kind of material, the current analogue media may have magnetic emulsion degradation over time, and this may be evident in the appearance of audible sound distortions that affect the intelligibility of the voice content or the quality of the music records. Other problems are the unavailability of player equipment for some of the media. The technological obsolescence that has accelerated in recent years forces new equipment to replace the old material, and requires maintenance of old players to read the old media format.
- Create a new online multi-user environment for access to the archive collection. The use of this collection required the search of the desired content through a documental application that might be accessible from any RNE LAN workstation, including powerful tools for the reference search. Nevertheless, once the content reference was identified, it was required to walk up to the central warehouse, fill out a lending request and take the material in-situ if it was already on loan to another user. In order to minimize this problem several copies of each document were stored which represents an additional cost. Now, once the content reference is identified it may be heard in preview mode or requested as a file download to the local PC and everything from the documentalist desk. Once it has been decided the audio file will be broadcast, it is exported directly to the Broadcasting application, without format change.
- Change the current RNE process for the incorporation of new entries. New applications have been created that allow the massive incorporation of material in digital format into the sound archive, as well as the digitization of new analogue entries. The metadata associated with each file are processed in the content catalogue in both cases, unifying the file formats ultimately stored in the archive.

4 Workflow management application

It was necessary to define a work process and workflow for the digitization process management. In addition, a new collaborative tool was developed based on Lotus Notes. Previous approaches were much less detailed and with less quality control than in RNE [8].

The process is described as follows:

- The sound archive documentalists select in the documental database the documents to be digitized during the week.
- A data exchange file is generated from the documental management legacy application that is read by the Lotus Notes application, where it is included the material data to be digitized and the associated documental data.
- The Lotus Notes application generates from this information different work orders. A work order is the work description to be made by a digitizer and includes the required information so the digitizer identifies the material to be digitized.
- The shift coordinator distributes the work orders between the 10 operators in each shift who receive the diary work description to be made at his desk.
- The operators reproduce the material monitoring the audio quality and the correspondence between the support and the given documental information.
- When the operators end the digitization, metadata are inserted, the sample speed is changed when it is required, and compressed and linear files are created.
- The files are sent to the central server and erased from the local workstation.
- The shift coordinator selects the material that is going to pass an additional quality control test and generates the quality work orders.
- The quality work orders are received by the two operators by shift that dedicate themselves exclusively to this work. Their mission is verifying the quality of the work already made, identifying the possible problems not detected by the digitization operator.
- Afterwards the RNE documentalists do additional quality control that is useful to certify the work effort.
- Once all this process is finished a data exchange file is generated that is sent to the legacy documental manager and informs him about the final status after the digitization and the real length once it has been digitized and blank spaces removed.

The Lotus Notes application can generate a wide variety of reports related to the digitized material, incidents and percentage of documents tested in the quality control. Additionally it has been statistically modeled the digitization process to get information about the quality of the process, the process non-identified error probability, and the margin of error of these measurements. For this purpose the Alcalá University developed a complex statistical study.

5 Quality Control data description

The results of the Quality Control processes have been periodically filed, guaranteeing a time period lower than a week. A number of 73 time periods have been considered, which cover digitalized files from May to December 2001.

The following values have been considered as data in each of the files provided by the quality control processes:

1. Number of digitized sound files in the considered period.
2. Number of tested files by IBM Quality Control.
3. Number of tested files by RNE Quality Control.
4. Errors detected in each quality control process, classified in accordance to the error class:
   - Jukebox Errors.
   - Human Errors.
   - Software Errors.
   - Documental Errors.
   - Support errors and other errors.

In these 73 time periods, 490,538 files have been digitalized, 106,725 files have been tested in the IBM Quality Control and 56,260 files in the RNE Quality Control.

6 Statistical model of one Quality Control

In this section, the model considered for each Quality Control stage considered in isolation, is described. The use of statistical models is quite common in the area of reliability or quality control [9][10].

Each digitalized file has been considered as a simple experiment. Each experiment is random with the following properties:

- Trials are independent.
- Each outcome is a sequence of trials, each one of which is a "success" or a "failure".
- The probability p of success and the probability q=1-p of failure is the same for each trial.

According to these properties, the experiment is binomial [11]. So, in our experiment a binomial random variable is sampled. The purpose is to

estimate the value of the probability p of success from these samples, which is a problem of parametric estimation. To solve it we apply the maximum likelihood method, obtaining the sample mean as optimum estimator with the considered criterion:

$$\hat{p} = \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} n_i} = \frac{\text{Success trials}}{\text{Total trials}} \tag{1}$$

Once the sampling distribution has been characterized, it is possible to estimate the confidence intervals in the calculation of the parameter p. It is necessary to estimate the value of the variance by looking for the maximum likelihood estimator. In a normal distribution the expression for which is:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} (p_i - \hat{p})}{N} \tag{2}$$

Theorem 1: From k independent N(0,1) random variables Z_i (i=1, 2, ..., k), a new random variable is formed from the sum of their squares $X = Z_1^2 + \Lambda + Z_k^2$. The random variable X has a chi-square distribution with k degrees of freedom (χ_k^2) and its pdf is:

$$f(x) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} x^{(k/2)-1} e^{-x/2} \quad , \quad x > 0 \tag{3}$$

Theorem 2: Given that Z and V, are two independent random variables, where Z is N(0,1) and V is chi-square distributed with k degrees of freedom (χ_k^2), then the random variable:

$$T = \frac{Z}{\sqrt{V/k}} \tag{4}$$

has a t-student distribution with k degrees of freedom (t_k) and its density function is:

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)} \frac{1}{\left[\left(x^2/k\right)+1\right]^{(k+1)/2}} \quad , \quad -\infty < x < \infty \tag{5}$$

Once variance estimation (2) and mean estimation (1) are known, a new random variable can be defined:

$$T = \frac{\hat{p}}{\sqrt{\hat{\sigma}^2/N}} \tag{6}$$

Keeping in mind theorems 1 and 2, *T* is a random variable with t-student distribution with N degrees of freedom.

Finally, the confidence interval (-∞, z) of the estimator $\hat{p}$ can be described. The confidence interval must be defined for a given confidence coefficient:

$$P\left[-\infty < \hat{p} < z\right] = \alpha \tag{7}$$

The constant α is the confidence coefficient of the estimate and (1-α) is the confidence level. Thus α is a measure of our confidence that the unknown parameter p is in the interval (-∞, z). If α is close to 1 we can expect with near certainty that this is true. Once known the pdf of the estimator $\hat{p}$, *z* value can be defined:

$$z = \hat{p} + t^{-1}(\alpha, N) \cdot \sqrt{\frac{\sigma^2}{N}} \tag{8}$$

where $t^{-1}(\alpha; N)$ is an inverse t-student distribution function with N degrees of freedom and α confidence.

7 Model of two serial Quality Control processes

On last subsection, error pdfs were considered for each Quality Control process independently. Both quality control processes are placed in serial on the digitalization process. Only a few data set is studied on each quality control. Corrupted files detected by IBM are not analyzed by RNE. So, the probabilities and the confidence intervals on this last stage do not allow to make inferences over the global digitalization process. Our goal is to estimate the error probability in the global digitalization process. For this purpose, it is necessary to model the digitalization process as a state chain.

Using the same random variable described on subsection 6, the statistical method is described:

1. N1 is the number of files obtained after the digitalization stage.
2. Each file is tested by the IBM Quality Control Process. The probability of a file being tested by this process is pIBM.
3. N1(1-pIBM) is the average number of files that are not tested by the IBM Quality Control.
4. The probability of error in a file tested in the IBM Quality Control Process is (1 - p1).

5. Taking into consideration that files with errors detected by IBM return at the start of the digitalization process, the mean number of files that can be analyzed by the RNE Quality Control is:

$$N_2 = N_1(1 - p_{IBM}) + N_1 p_{IBM} \cdot p_1 \quad (9)$$

6. So, if the probability of error of a file being tested by the RNE Quality Control Process is pRNE , N2·pRNE files are the mean number of files tested by the RNE Quality Control. So, the mean number of files that are not tested by any quality control process is:

$$N_{not\_tested} = N_1(1 - p_{IBM})(1 - p_{RNE}) \quad (10)$$

7. The mean number of tested files is given by:

$$N_{tested} = N_1 p_{IBM} + N_1(1 - p_{IBM}) p_{RNE} \quad (11)$$

8. The error probability on the input files is determined by:

$$p_{error} = \frac{error\_number}{N_{tested}} \quad (12)$$

This expression is assumed to be a good estimation of the probability error in the digitalization process.

Once the error probability is calculated, the residual corrupted files after the digitalization process are determined by multiplying the estimated error probability by (1-pIBM) and (1-pRNE), y by the number of files. So, we can translate the estimated error probabilities and the estimated confidence intervals to the mean number of residual corrupted files, and the maximum number of errors for a given confidence probability.

Table 1: Estimated probabilities for the digitalization process between May and December, 2001

Residual Corrupted File Maximum Error Probability							
Error class		*Jukebox*	*Human*	*Soft*	*Doc.*	*Other*	*Total*
Mean Error probability		0,04%	0,08%	0,03%	0,01%	0,01%	**0,20%**
Max. Prob. for a given Confidence	90%	0,06%	0,09%	0,05%	0,01%	0,23%	**0,23%**
	99%	0,07%	0,10%	0,07%	0,02%	0,26%	**0,25%**
	99,9%	0,08%	0,11%	0,08%	0,02%	0,27%	**0,27%**
	99,99%	0,09%	0,12%	0,09%	0,02%	0,28%	**0,28%**
	99,9999%	0,10%	0,13%	0,10%	0,02%	0,31%	**0,31%**

8 Quality Control Results

Table 1 shows results obtained applying the statistical model to the state chain. Due to the low value of the estimated error probabilities, the Sound Archive Digitalization Process, on the analyzed dates, can be considered reliable. Over the 490.538 digitalized files we can asses with a confidence probability of 99.99% that less than 1.396 files are corrupted, which is a satisfactory result.

9 Conclusions

The RNE Archive digitisation process has been described in this paper. Special attention has been paid in describing the statistical model of the Quality Control introduced. This statistical model allows to estimate the probability of error in the digitisation process for a given
confidence. Several kinds of errors have been considered, like human errors, mechanical errors, software errors or support errors.
Tools have been developed to achieve the project objectives. IBM Global Services and RNE have fulfilled the project on time. RNE is very satisfied with the new online service they are providing to their internal users, and perhaps in the future to external users. There are no insurmountable obstacles to tackle these kinds of projects. It is just a matter of creating a knowledgeable working team, planning the project suitably, analysing the contents and executing with quality. We have now accomplished the objective of preserving our audio heritage. Furthermore, the results provided by the statistical model used to estimate the probability of error in the digitisation of sound material show that the reliability of the process is very high.

References

[1] UNESCO: *Text of the draft UNESCO Resolution on Digital Preservation. Proposed by the Conference of Directors of National Libraries (CDNL)*. The Hague, June 2001.

[2] Schuller, D.: Preserving Audio and Video Recordings in the Long-term. *International Preservation News*, 14 (1997).

[3] Schuller, D.: Preserving the Facts for the Future: Principles and Practices for the Transfer of Analog Audio Documents into the Digital Domain. *Journal of the Audio Engineering Society*, 49 (2001), 7/8, 618-621.

[4] Hafner, A.: The Suedwestrundfunk (SWR) and the Mass Storage Systems in Its Radio Sound Archives: Concepts and some Performance/Cost Aspects. *106th Audio Engineering Society Convention*. Munich, Germany, May 08-11 1999.

[5] IASA-TC 03. *The Safeguarding of the Audio Heritage: Ethics, Principles and Preservation Strategy. Version 2*, September 2001.

[6] Herla, S., Houpert J. and Lott, F.: From Single-Carrier Sound Archive to BWF Online Archive – A New Optimized Workstation Concept. *Journal of the Audio Engineering Society*, 49 (2001), 7/8, 606-617.

[7] Wright, R.: Broadcast Archives: Preserving the future. *BBC Information & Archives* (2001).

[8] Viering, H, Koide N. and Gereon, M.: High-Speed Analogue Tape to Digital File Conversion for Broadcast Libraries. *106th Audio Engineering Society Convention*. Munich, Germany, May 08-11 1999.

[9] Shridharbhai, K.: *Probability and Statistics with Reliability, Queueing, and Computer Science Applications*, 2nd Edition, John Wiley & Sons, 2001.

[10] Blischke, W.R., Prabhakar, D.N.: *Reliability: Modeling, Prediction, and Optimization*, Wiley-Interscience, 2000.

[11] Papoulis, A.: *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill International Editions, 1991.