GHENT UNIVERSITY

Applications of duplicate detection in music archives: from metadata comparison to storage optimisation.

The case of the Belgian Royal Museum for Central Africa

Joren Six, **Federica Bressan**, Marc Leman

IPEM, Ghent University, Belgium

IRCDL 2018 - 26 January 2018 - Udine, Italy

Overview I

Duplicate detection

- Applications for duplicate detection
    - To complete meta-data
    - To improve listening experiences
    - To segment tracks
    - To merge archives
- Robustness against speed changes

Acoustic fingerprinting

Case studies

- Case study: RMCA archive
- Case study: IPEM archive

Conclusion

Duplicate detection

Definition (Duplicate detection system)

A system that is able to compare every audio fragment in a set with all other audio in the set to determine if the fragment is **either unique or appears multiple times** in the complete set. The comparison should be **robust** against various artefacts.

Duplicate detection

Duplicates contain the *same recorded event* but can differ by:

- Noise from various sources
  - Carrier dependent
    - Magnetic tape hum/hiss
    - Phonographic disc pop/clicks...
  - Imperfections from A/A or A/D conversion, among which changes in playback speed
- Various dynamics artefacts: intensity, compression, ...
- Digital encoding format

GHENT UNIVERSITY

Duplicate detection to complete meta-data

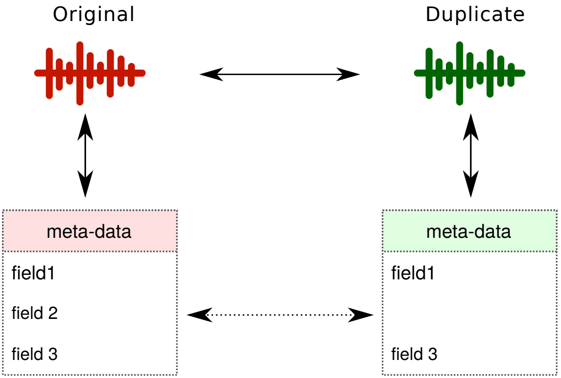

Figure: Duplicate detection to complete meta-data.

Duplicate detection to improve the listening experience

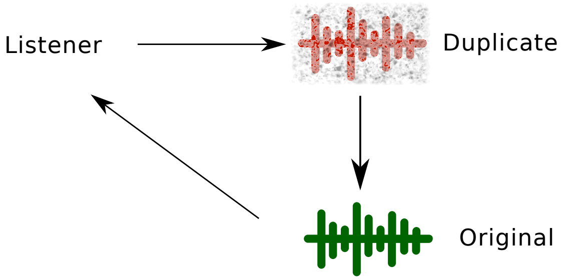

Figure: Duplicate detection to improve the listening experience.

Duplicate detection for segmentation

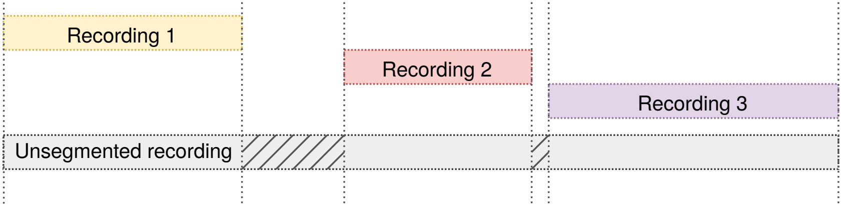

Figure: Duplicate detection for segmentation.

Duplicate detection for merging archives

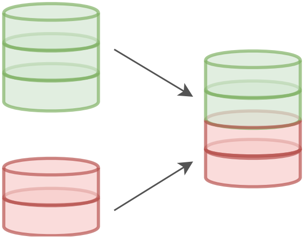

Figure: Merging two archives: two plus three equals four.

Allows to identify *unique items* in merged archives. All above applications apply

- Meta-data improvement
- Improved listening experience
- Reuse segmentation points

Robustness against speed changes

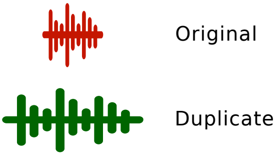

Figure: Robustness against speed changes.

Robustness to speed change is needed if:

- Many wax cylinders are present
- Uncalibrated tape recorders were used
- For historical archives consisting of merged archives

Acoustic fingerprinting

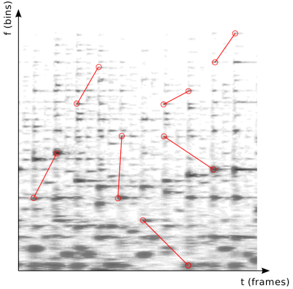

Figure: An acoustic fingerprinting approach

- Mature MIR technology
- Allows duplicate detection
- Efficient algorithms [5, 1, 3]
- Some robust to speed change [3, 4]
- Implementations available [3]

Acoustic fingerprinting

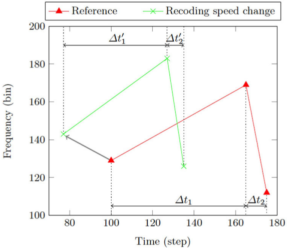

Figure: The effect of speed modification on a fingerprint

The software used is Panako:

Article	Panako [3]
Website	http://panako.be
License	GNU Affero GPL

To operate Panako you do not need an MIR specialist

GHENT UNIVERSITY

Case study: RMCA archive

Figure: Meta-data on file at the RMCA-archive

Collection of the Royal Museum for Central Africa, Tervuren, Belgium [2]

- More than 35 000 items
- Mainly field recordings from Central Africa
- First recordings from 1890s
- Many analogue carriers types
- Challenging meta-data

GHENT UNIVERSITY

Case study: RMCA archive

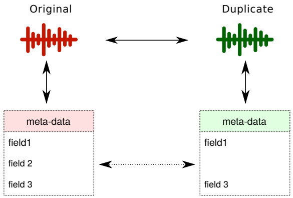

Figure: Main application: segmentation re-use

Duplicate detection on this large historical archive has to aims:

- Compare meta-data between pairs
- Quantify the amount of duplicates

2.5% (887 of 35306) recordings were found to be duplicates

RMCA archive

Field	Empty	Different	Exact match	Fuzzy or exact match
Year	20.83%	13.29%	65.88%	65.88%
People	21.17%	17.34%	61.49%	64.86%
Country	0.79%	3.15%	96.06%	96.06%
Province	55.52%	5.63%	38.85%	38.85%
Place	33.45%	16.67%	49.89%	55.86%
Language	42.34%	8.45%	49.21%	55.74%
Title	42.23%	38.40%	19.37%	30.18%
Collector	10.59%	14.08%	75.34%	86.71%

Table: Comparison of pairs of meta-data fields

RMCA archive

Original title	**Duplicate title**
Warrior dance	Warriors dance
Amangbetu Olia	Amangbetu olya
Coming out of walekele	Walekele coming out
Nantoo	Yakubu Nantoo
O ho yi yee yi yee	O ho yi yee yie yee
Enjoy life	Gently enjoy life
Eshidi	Eshidi (man's name)
Green Sahel	The green Sahel
Ngolo kele	Ngolokole

Table: Pairs of fuzzy matching titles.

Case study: IPEM archive

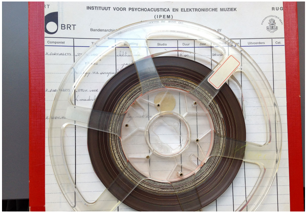

Figure: Open-reel tape from the IPEM archive

The archive of Institute for Psychoacoustics and Electronic Music (IPEM)

- About 1800 open reel tapes
- Early electronic music
- Represent 1960s-1970s musical avangarde in Belgium

Case study: IPEM archive

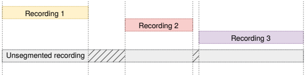

Figure: Main application: segmentation reuse

The archive has been digitized twice. Once in 2001 and in 2014 with higher quality. Planned to re-use segmentation and meta-data from first digitization.

Conclusion

- Presented applications of duplicate detection
- Acoustic Fingerprinting allows duplicate detection
- Illustrated applications with two case studies
- Pointer to software for duplicate detection

GHENT UNIVERSITY

Bibliography I

Jaap Haitsma and Ton Kalker.
A highly robust audio fingerprinting system.
In *Proceedings of the 3th International Symposium on Music Information Retrieval (ISMIR 2002)*, 2002.

Joren Six, Federica Bressan, and Marc Leman.
Applications of duplicate detection in music archives: From metadata comparison to storage optimisation - The case of the Belgian Royal Museum for Central Africa.
In *Proceedings of the 13th Italian Research Conference on Digital Libraries (IRCDL 2018)*, In Press - 2018.

Bibliography II

Joren Six and Marc Leman.
Panako - A scalable acoustic fingerprinting system handling time-scale and pitch modification.
In *Proceedings of the 15th ISMIR Conference (ISMIR 2014)*, pages 1–6, 2014.

R. Sonnleitner and G. Widmer.
Robust quad-based audio fingerprinting.
Audio, Speech, and Language Processing, IEEE/ACM Transactions on, PP(99):1–1, 2016.

Bibliography III

Avery Li-Chun Wang.
An industrial-strength audio search algorithm.
In *Proceedings of the 4th International Symposium on Music Information Retrieval (ISMIR 2003)*, pages 7–13, 2003.