An audio-only method for advertisement detection in broadcast television content

António Ramires, Diogo Cocharro, Matthew Davies

INESC TEC - Sound and Music Computing Group
antonio.ramires@inesctec.pt

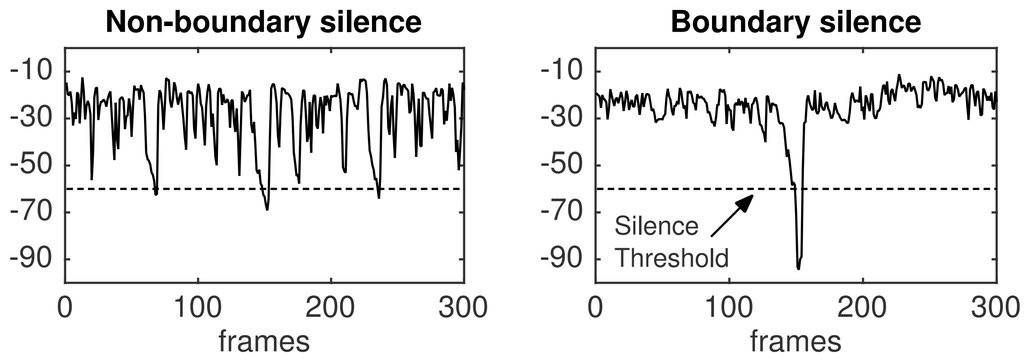
Introduction

- The classification of audiovisual content into categories and the identification of advertisement has become increasingly important for end-users, broadcasters and stakeholders. This has special importance in the case of television content both for the need to archive content without advertisement, and in streaming contexts to allow for the region-specific substitution of advertising.
- Currently, this process is typically performed by a human operator. As a result, the process is labour-intensive and expensive.
- While typically approached from a video-only or audio-visual perspective, **we present an audio-only method**, which centres on the detection of short silences which exist at the boundaries between programming and advertising, as well as between the advertisements themselves.

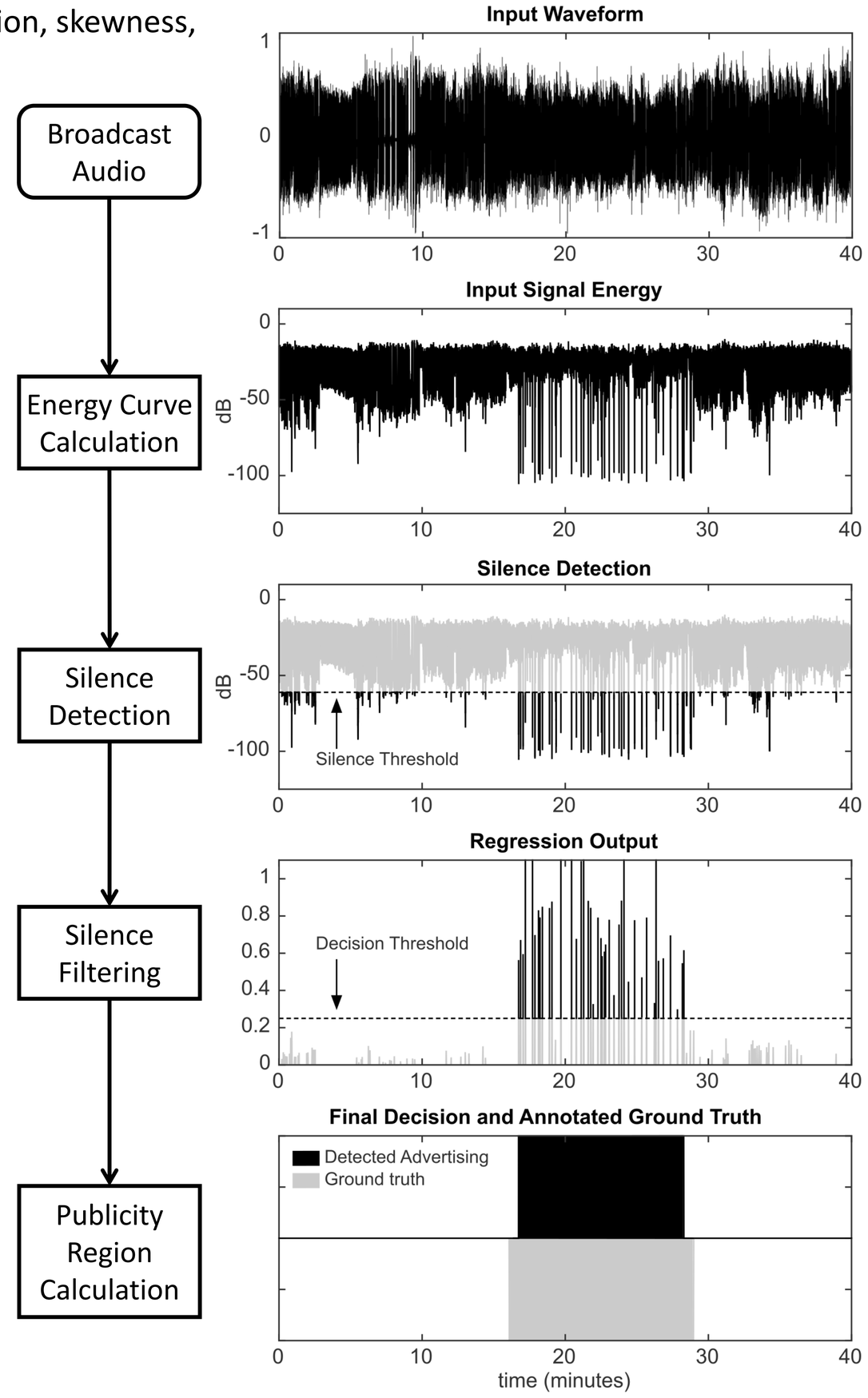
Methodology

Our approach centres on the existence and detection of short pauses of silence (i.e., very low audio signal energy) in between separate pieces of content. We now provide an overview of each stage of the algorithm:

1) The audio is partitioned into non-overlapping frames which are synchronized with the video frame rate (25 fps)
2) For each frame, we calculate the signal energy, in dB, to force all low energy parts of the signal to take large negative values.
3) To find the low energy points in the input signal, we compare e_i at each frame, i to a silence threshold, $\eta = 60dB$, and retain those frames for which $e_i \leq \eta$.
4) Since short regions of silence can occur naturally within programming, we must filter them. A **boundary silence** is considered short in duration, with a low minimum value, surrounded with higher energy.
   We collect the max, mean, min, inter-quartile range, standard deviation, skewness, and kurtosis from a local context window of 6s around each silence.



5) Multiple linear regression on the extracted features is performed and all detected silences greater than the decision threshold, $\beta = 0.25$ are retained and set to a value of 1.
6) We slide a rectangular window of 150s across the output and consider a region of advertising those which:
   - The window contains more than 1 silence.
   - The advertisement region is longer than 60s.
7) The detected advertising region starts at the frame where the first silence exits the long-term window. Likewise the region ends at the frame when the final silence exits the long-term window.



Results

Input Channel	Total Duration	Advertising Duration	ComSkip* Accuracy	Proposed Alg. Accuracy
RTP 1_a	6h52m	0h23m	0.426	0.782
RTP 1_b	1h10m	0h11m	0.007	0.863
RTP 2	8h25m	0h24m	0.366	0.499
SIC	8h37m	2h18m	0.648	0.931
TVI	1h13m	0h9m	0.794	0.966
Overall	26h17m	3h24m	0.610	**0.874**

**ComSkip is a freely available commercial detector, available at https://github.com/erikkaashoek/Comskip*

This article is a result of the project MOG CLOUD SETUP – №17561, supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF).
Diogo Cocharro was supported by Project TEC4Growth-Pervasive Intelligence, Enhancers and Proofs of Concept with Industrial Impact/NORTE-01-0145-FEDER-000020 is financed by the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, and through the European Regional Development Fund (ERDF).