A ZOOM FILTER FOR APPLAUSE AND LAUGHTER

Research Review

Structure

1. Corpora
2. Laughter Detection
3. Applause Detection
4. Next Steps

CORPORA

Corpora available

- AudioSet by Google - 2017
  - 5.8 thousand hours of audio
- ICSI meetings database - 2004
  - 70h of transcribed meeting data
- Switchboard - 1997
  - 260 hours of transcribed telephone conversations
- SSPNet-Mobile Corpus - 2014
  - 12hours of annotated telephone conversations

include laughter
but **not** applause

Chosen Corpus

- **AudioSet** by Google
  - 2.1 million annotated videos
  - 5.8 thousand hours of audio
  - 527 classes
  - 2,084,320 labeled 10s snippets

- **Applause**
  - 2247 videos - 6.2 hours
  - est. accuracy: 90% (9/10)
  - Human sounds > Human group actions
  - Sub-categories:
    - None

- **Laughter**
  - 5696 videos - 15.8 hours
  - est. accuracy: 100% (10/10)
  - Human sounds > Human voice > Laughter
  - Sub-categories:
    - Baby laughter, Giggle, Snicker, Belly laugh, Chuckle/chortle

Temporally-Strong Labels

(May 2021)

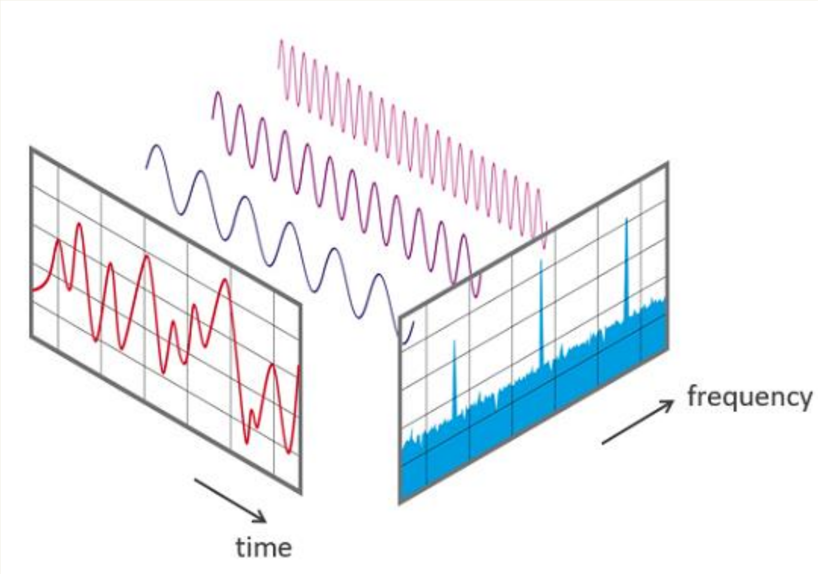
"The Benefit of Temporally-Strong Labels in Audio Event Classification" - Hershey et al

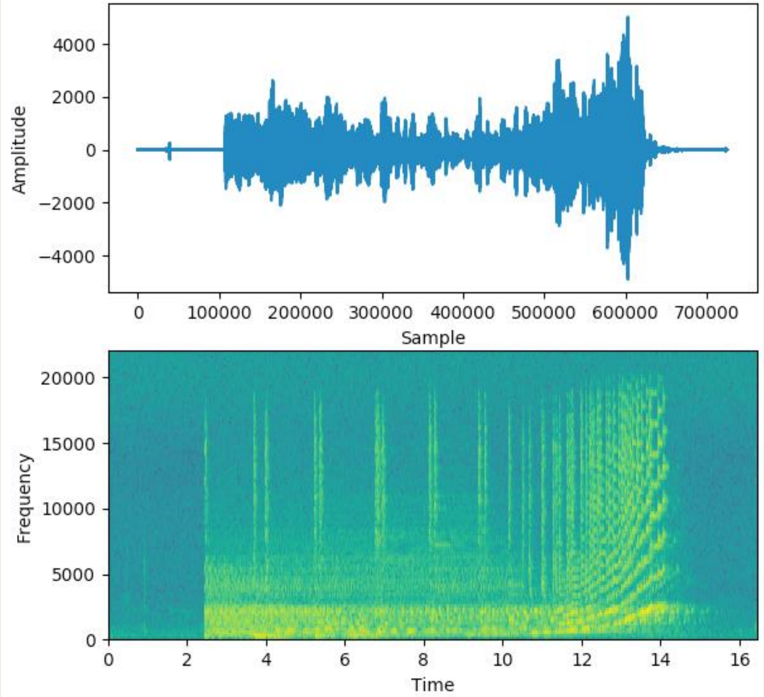
- Updated version of AudioSet with:
  - manual boundary setting -> clip lengths
  - more accurate descriptions
  - considers multiple occurrences in one 10s clip
- Paper results state
  - that a classifier trained on the large 'weakly-labeled' dataset can be improved via-fine-tuning on 'strongly-labeled' data

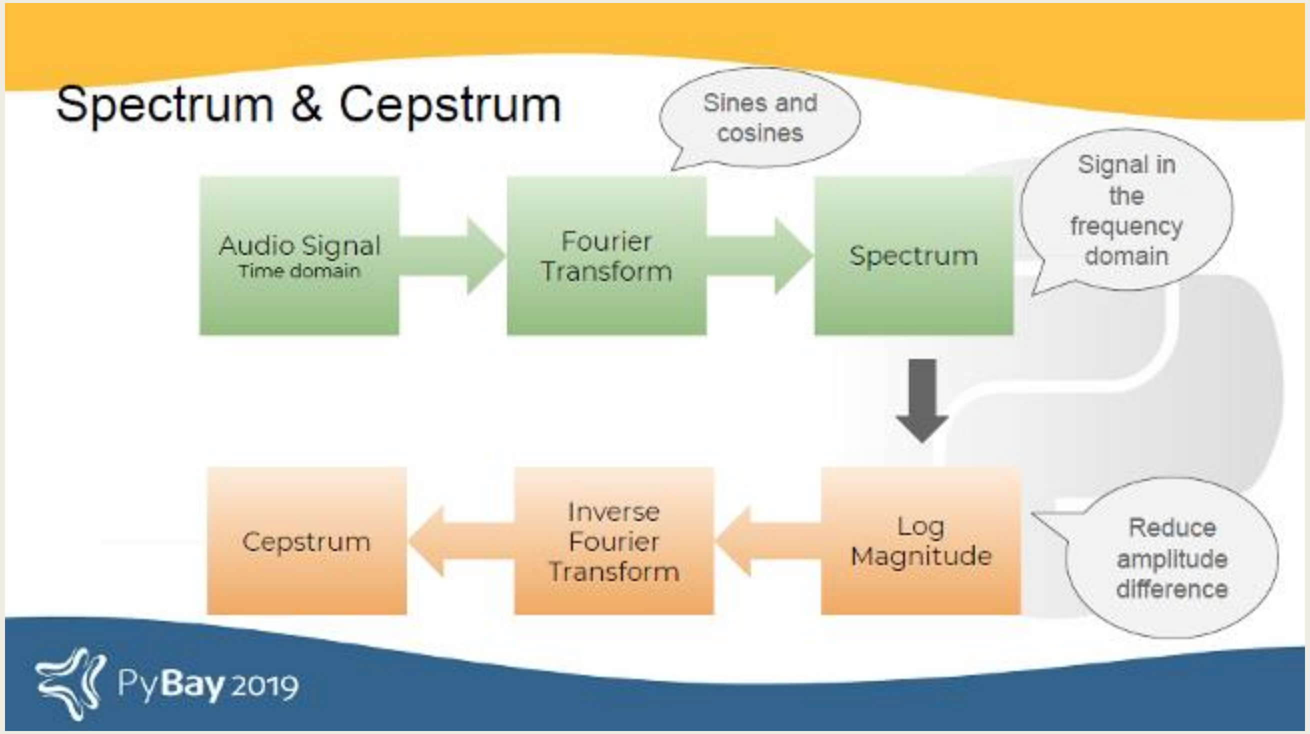
LAUGHTER DETECTION

Detection using different data types

- audio-visual, visual or sensor data
  - Turker, B. B., Yemez, Y., Sezgin, T. M., & Erzin, E. (2017). **Audio-facial laughter detection in naturalistic dyadic conversations.** *IEEE Transactions on Affective Computing*, *8*(4), 534-545.
  - Akhtar, Z., Bedoya, S., & Falk, T. H. (2018, April). **Improved Audio-Visual Laughter Detection Via Multi-Scale Multi-Resolution Image Texture Features and Classifier Fusion.** In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 3106-3110). IEEE.
  - Hagerer, G., Cummins, N., Eyben, F., & Schuller, B. (2018, April). **Robust laughter detection for wearable wellbeing sensing.** In *Proceedings of the 2018 International Conference on Digital Health* (pp. 156-157).







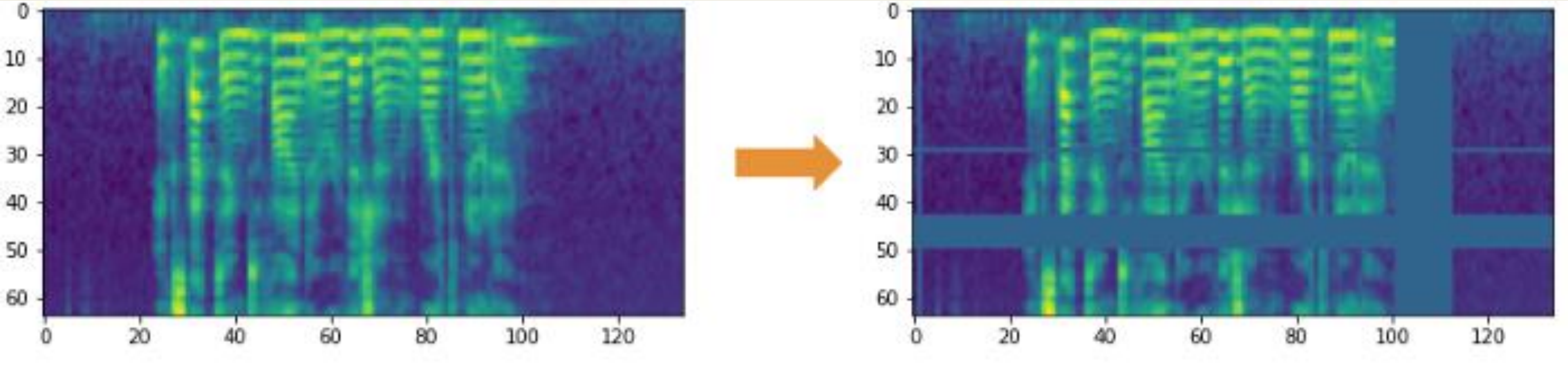
sources:

- https://towardsdatascience.com/audio-deep-learning-made-simple-part-1-state-of-the-art-techniques-da1d3dff2504
- https://opensource.com/article/19/9/audio-processing-machine-learning-python#comments

Robust Laughter Detection in Noisy Environments
(Gillick et al. - Sept. 2021)

- States that prior work performs badly in noisy environments
- Uses Switchboard dataset (SLD) and AudioSet (WLD)
- New annotations for part of AudioSet for evaluation
  - precise segmentations - start and end points of each laugh
    - 148min of audio – 1000 laughter snippets
      - *58min laughter – 1492 distinct laughter events*
    - Additional 1000 clips without laughter for testing
      - 20% laughter in testing set
- Compares three models
  1. baseline feed-forwards NN with engineered features
  2. ResNet model on spectrogram data
  3. ResNet model augmented with data transformations

- Baseline
  - NN on top of traditional audio features like MFCC's (Mel Frequency Cepstral Coefficients)
- ResNet – Residual NN
  - features learned from spectrogram
- ResNet with Data Augmentation
  - Adding noise, masking spectrogram sections, pitch-shifting, time-stretching



Results

Table 2: *Laughter detection performance on Switchboard and AudioSet, with 95% bootstrap confidence intervals. Segment-based metrics are reported per frame for Precision, Recall, and F1 scores. SLD and WLD refer to strongly and weakly labeled data.*

Train on Switchboard (SLD)	Results on Switchboard Test Data			Results on AudioSet Test Data		
	PRECISION	RECALL	F1	PRECISION	RECALL	F1
Baseline	0.634 (±0.025)	0.752 (±0.023)	0.688 (±0.016)	0.224 (±0.016)	0.901 (±0.014)	0.359 (±0.021)
ResNet	0.677 (±0.022)	0.830 (±0.019)	0.747 (±0.017)	0.464 (±0.020)	0.748 (±0.018)	0.573 (±0.018)
ResNet + Augmentation	0.676 (±0.022)	0.847 (±0.018)	**0.752** (±0.016)	0.508 (±0.020)	0.759 (±0.017)	**0.608** (±0.015)
Train on AudioSet (WLD)						
Baseline	0.300 (±0.024)	0.765 (±0.026)	0.430 (±0.026)	0.372 (±0.019)	0.856 (±0.019)	0.519 (±0.019)
ResNet	0.439 (±0.036)	0.710 (±0.028)	0.542 (±0.030)	0.371 (±0.017)	0.928 (±0.012)	0.530 (±0.018)
ResNet + Augmentation	0.468 (±0.027)	0.700 (±0.025)	0.563 (±0.023)	0.385 (±0.018)	0.925 (±0.015)	0.545 (±0.018)

- Evaluation on individual frames lasting 23 ms

APPLAUSE DETECTION

Applause Sound Detection
(Christian Uhle, 2011)

- Features: combination of MFCC and LLD (low-level descriptors)
- Uses MLP and SVM with radial basis function for classification
- Uses only 210 snippets (of 9-30s length)
  - *90/10 split - means only 21 test samples*
- Claims 'real-time detection with low-latency'
- Results:

	Predicted Applause	Predicted No Applause
Applause	83%	2%
No Applause	3%	12%

Characteristics-based effective applause detection for meeting speech
(Li et al. , 2009)

- Uses a 4-layer decision tree
- Unpublished dataset:
  - 50h of multi-participant meeting speech
  - *500 appluase segments (0.8-36s)*

Table 2
Comparisons between the proposed algorithm and the traditional algorithm.

Parameters	The proposed algorithm	The traditional algorithm
PR (%)	94.34	91
RR (%)	98.04	94.12
*F*1-measure (%)	96.15	92.53
Computational time (min)	59.4	92.5

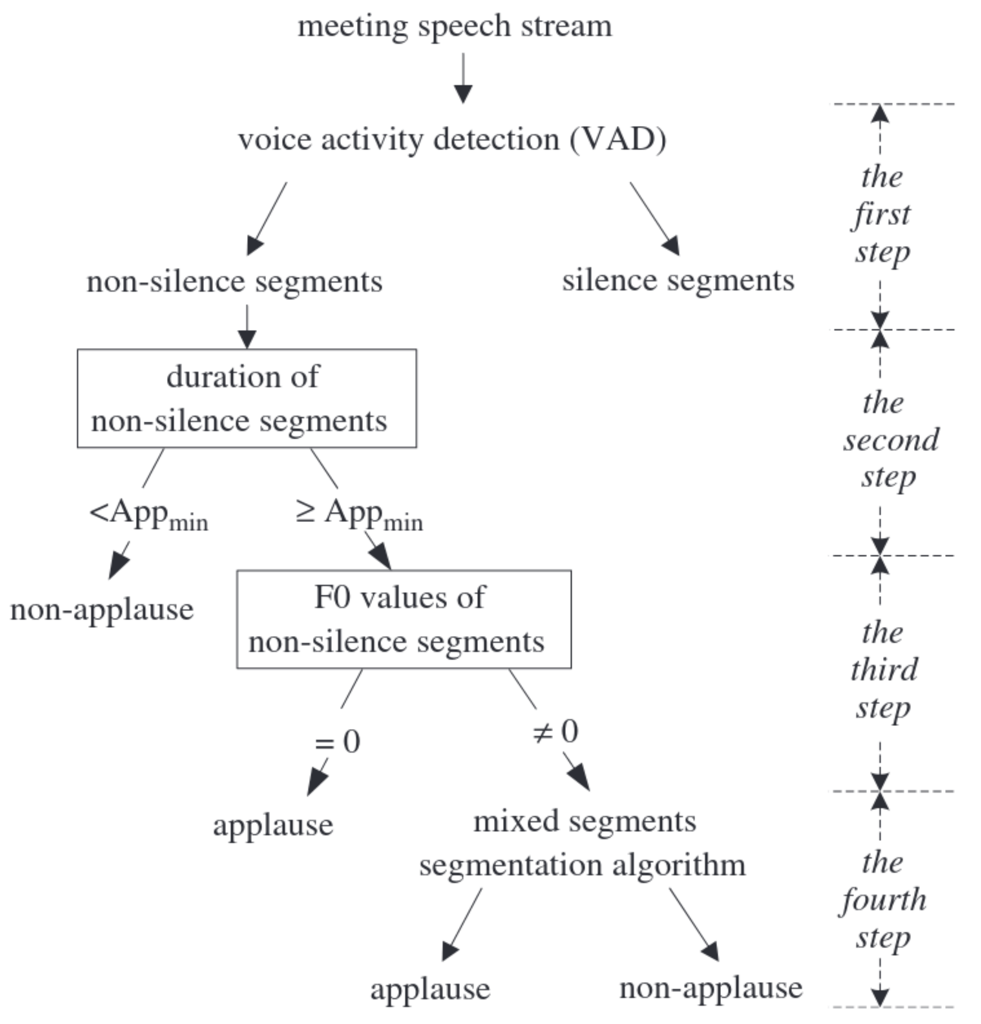


Fig. 3. The framework of the proposed algorithm.

- Novel approach for detecting applause in continuous meeting speech
  (Manoj et al., 2011)

- decision tree structure with 4 thresholds
  1. Silence detection
  2. Energy decay factor - distinguish speech and noise
  3. First Local Minimum of autocorrelation function – also distinguishes noise and applause
  4. Band Energy ratio – again distinguish applause from speech and silence
- Results
  - Conventional method
    - 36-dimensional feature vector (from MFCCs) fed into GMM (Gaussian Mixture Model)

method	Precision rate	Recall rate	F1 score
proposed	94.40%	90.75%	92.54%
conventional	67.47%	96.13%	79.29%

Next Steps

- Use the AudioSet Corpus
  - Utilising the extra annotations from Gillick et al.
  - As well as the temporally-strong-labeled subset
- Challenge
  - Combine the corpus data mentioned above
  - Combine the algorithms from laughter and applause
    - Applause: suggests a decision tree approach
    - Laughter: suggests a NN approach