USC University of Southern California

USC Viterbi School of Engineering

Ming Hsieh Department of Electrical Engineering

Paralinguistic Event Detection from Speech using Probabilistic Time Series Smoothing and Masking

Rahul Gupta, Kartik Audhkhasi, Sungbok Lee, Shrikanth Narayanan

Signal Analysis and Interpretation Lab (SAIL)

Introduction

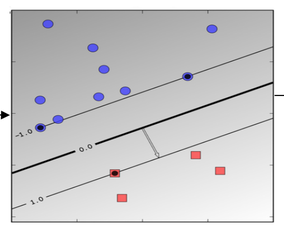
Frame-wise detection of laughters and fillers in telephonic speech

Database: SSPNet Vocalization Corpus
- 2763 audio clips
  - 1583 train, 500 dev, 680 test
  - Each clip 11 seconds long
  - Each clip has at least one vocal event (laughter or filer)

Detection System

- Baseline system

141 Frame – wise features (ex: MFCC+Δ+ΔΔ, F0+Δ+ΔΔ etc.) @ 100 frames/sec



Multi-class SVM Classifier

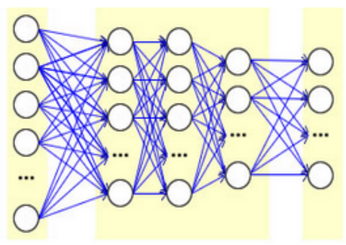
Frame level Decisions @ 100 Frames/sec

$P_1(E),P_2(E) \dots P_{1100}(E)$

E ϵ {Laughter, Filler}

- Applied system

141 Frame – wise features (ex: MFCC+Δ+ΔΔ, F0+Δ+ΔΔ etc.) @ 100 frames/sec



Deep Neural Network

$P_1(E),P_2(E) \dots P_{1100}(E)$

Smoothing

Masking

Frame level Decisions @ 100 Frames/sec

Deep Neural Network

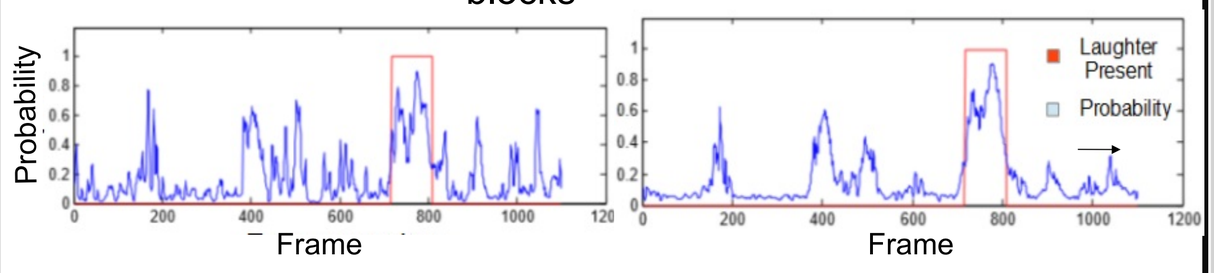
- Better Discriminative model
  - Not necessarily linear boundary as SVM
  - Recent emergence as a strong classifier scheme*

- 4 layers
  - (141 x 47 x 47 x 3) configuration
  - Sigmoidal Activation function

Smoothing

Smooth out time series temporally

$P_1(E),P_2(E) \dots P_{1100}(E)$ → Low Pass Filter H(z) → Delay / Advance blocks (Z^a_1, Z^a_2, Z^a_k) → Logistic Regression → Smoothed Probabilities

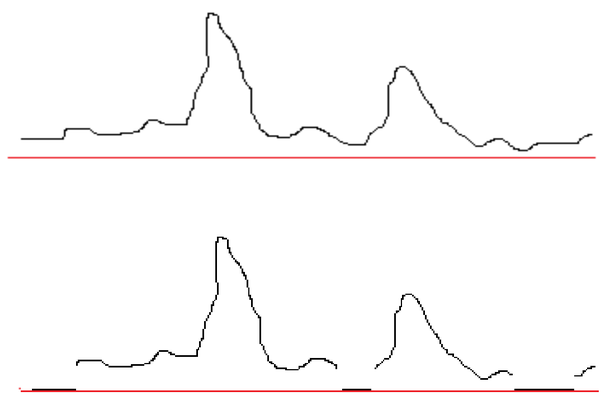


Before smoothing

After smoothing

Masking

Mask the time series to reduce false alarms based on
- Time series properties
- Entropy of output probabilities from an ASR system



- Attenuate probability around low probability Frames

- Increase probability around high probability frames

Results and Future work

System	AUC		UWAUC
	Laughter	Filler	
Baseline System	86.2%	89%	87.6
DNN	90.1%	90.1%	90.1%
Smoothing	94.6%	94.4%	94.5%
Masking	95.1%	94.7%	94.9%

Future work:

- Improve classification system Use better schemes to capture temporal context More features, better classification …
- Use the system For predicting higher level behavioral attributes Emotion Distress Engagement etc..

Ming Hsieh Institute

Ming Hsieh Department of Electrical Engineering