Statistical Language and Speech Processing, Seventh International Conference, SLSP 2019
Ljubljana, Slovenia - October 14-16, 2019

Emotion and Behavioral Tracking in the Lab and in the Wild

Alexandros Potamianos
Associate Professor, National Technical University of Athens
Chief Technology Officer and co-Founder, Behavioral Signals Technologies
Adjunct Associate Professor, University of Southern California

Contributors

- Athanasios Katsamanis, *Behavioral Signals*
- Theodoros Giannakopoulos, *Behavioral Signals*
- Aggelina Chatziagapi, *Behavioral Signals*
- Efthymios Georgiou, *NTUA*, *Behavioral Signals*
- Efthimios Tzinis, *U. Illinois at Urbana-Champaign*
- George Paraskevopoulos, *NTUA, Amazon*
- Shrikanth Narayanan, *USC*, *Behavioral Signals*
- George Pantazopoulos, *Behavioral Signals*
- Dimitris Sgouropoulos, *Behavioral Signals*
- Malvina Nikandrou, *Behavioral Signals*
- Spiridon Dimopoulos, *Behavioral Signals*
- Nikolaos Ellinas, *NTUA*, *Innoetics/Samsung*
- Christos Baziotis, *U. of Edinburgh*
- Jimmy Gibson, *Behavioral Signals*
- Colin Vaz, *Behavioral Signals*
- Tanner Sorensen, *USC*, *Behavioral Signals*
- Spiros Georgiladakis, *Behavioral Signals*
- Vicki Kolovou, *Behavioral Signals*
- Dogan Can, *Apple Research*

BEHAVIORAL SIGNALS

Outline

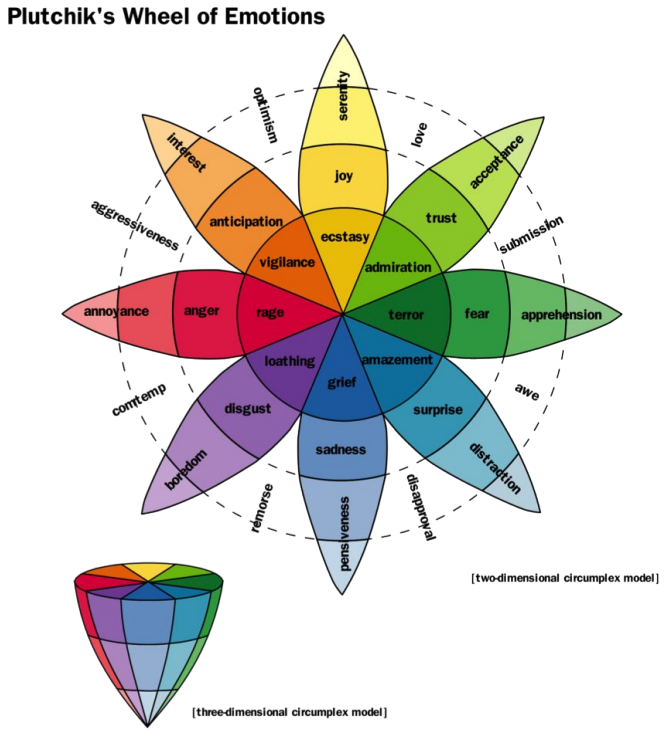
- Problem Definition
- Baseline Emotion Tracking System
  - Features
  - Machine Learning Architectures
- Advanced Feature Extraction
- Data Sparsity
  - Data Selection - Active Learning
  - Data Augmentation
- Multimodal Fusion
- In the Lab vs in the Wild:
  - Baselining a Speaker
  - Tuning your Operation Point
  - Context
  - Emotional Dynamics
- Applications and Demos

Problem Definition

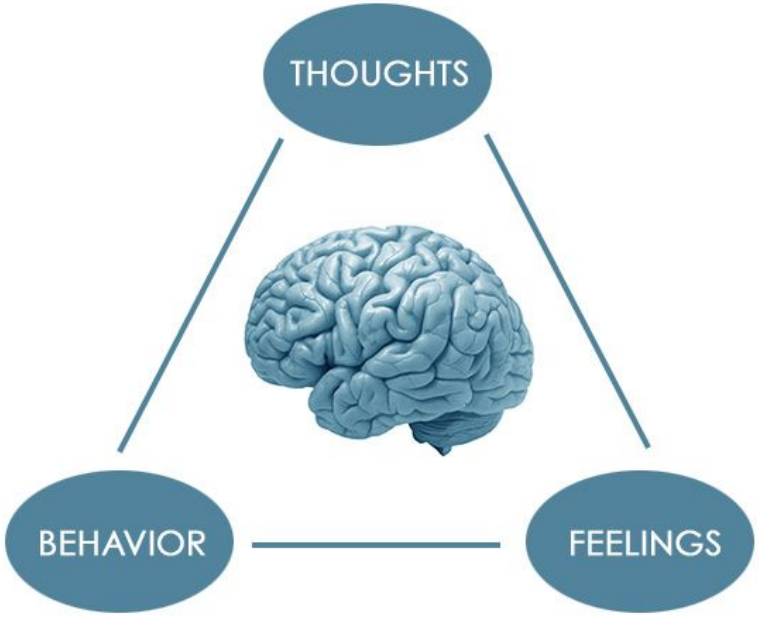
- Emotions as Behavioral Motivators
- The triptych of emotion, thoughts and behaviors
- Behavioral Signals

Definition of Emotions

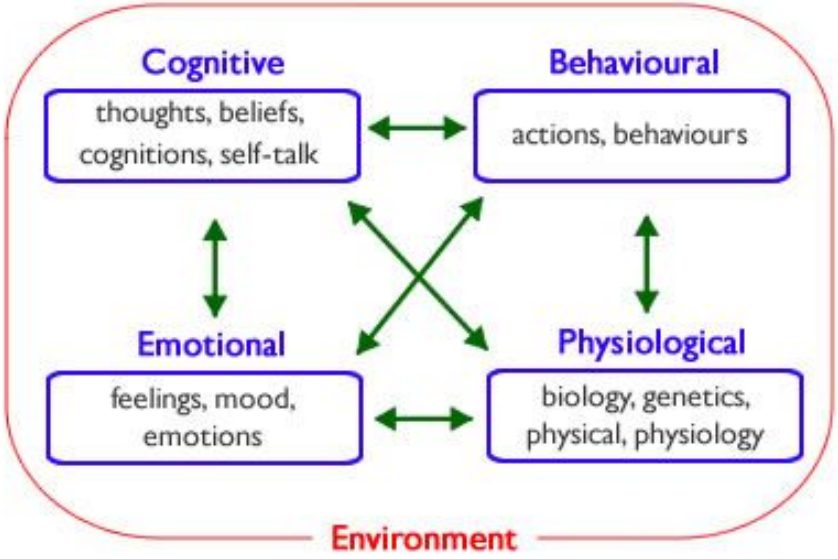
- Psychology:
  - Many theories interpret and categorize emotions
- Emotions as Behavioral Motivators
- Continuous vs Discrete emotions
- Primary vs Secondary Emotions
  - Judging events vs predictions/thoughts
- System 1 vs System 2
- Regulation of Emotions



Emotions, Behaviors and Thoughts



Cognitive Behavioral Therapy



Behavioral Signals: What is that?

"... **behavioral signals are the observable part of a behavior,** i.e., how this behavior is expressed, perceived and captured multimodally by humans (and possibly machines). Behavioral signals are slices of information — including those characterized as thin slices— that need to be combined and composed to jointly describe a behavioral event, action or state ..."

https://medium.com/behavioral-signals-ai/behavioral-signals-what-is-that-367ba0de49d2

BEHAVIORAL SIGNALS

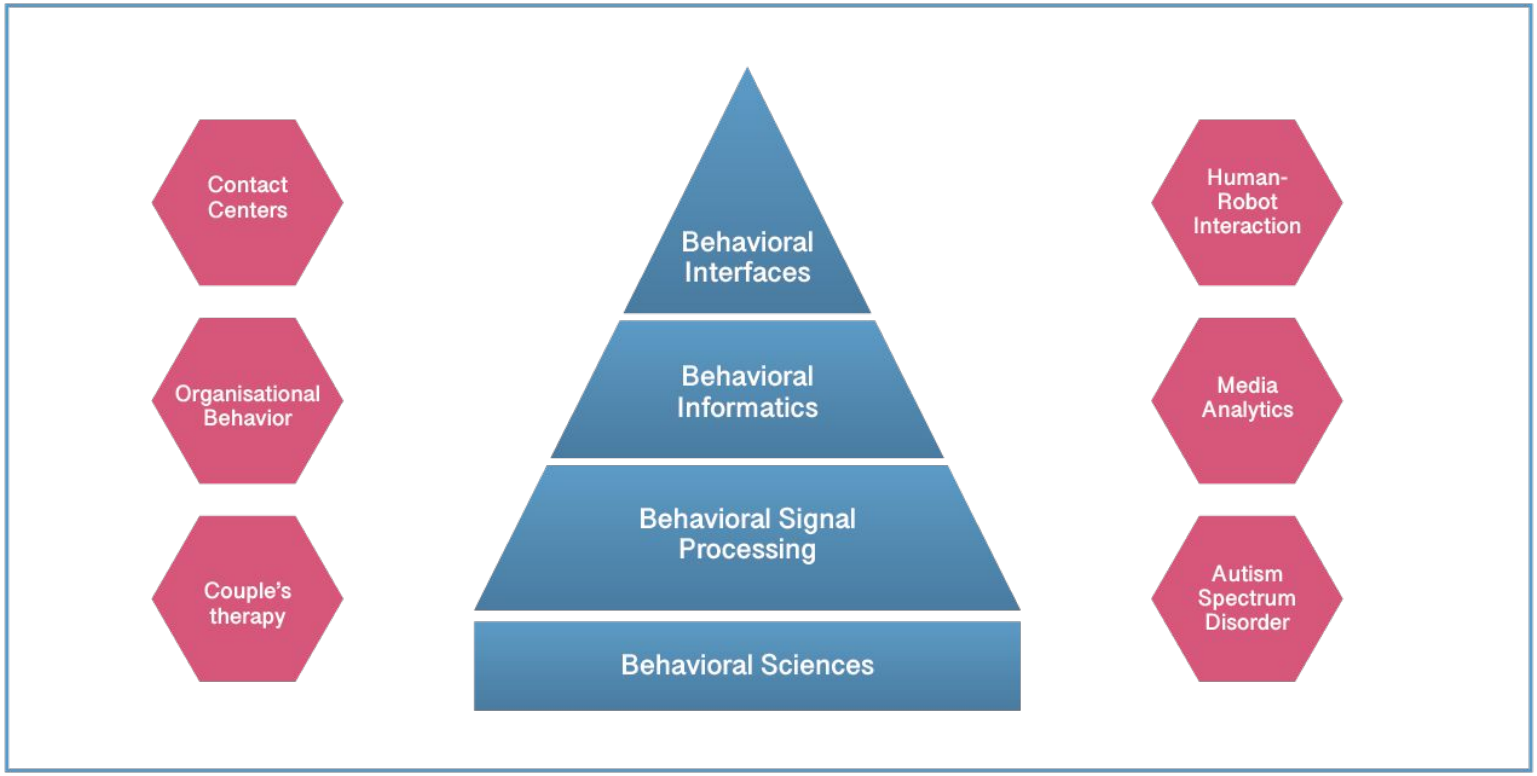
What drives our behavior

"Roughly speaking our behaviors are driven or conditioned by **who we are** (personality, morality), **how we feel** (emotion, mood), **how we are perceived** and think of ourselves **socially** (status, dominance, rapport), **what are our goals** and how we plan to achieve them (persuasion), **what we believe in**, and — to make things even more complex — **how self-aware we are** about all these things (awareness, regulation, attention)"

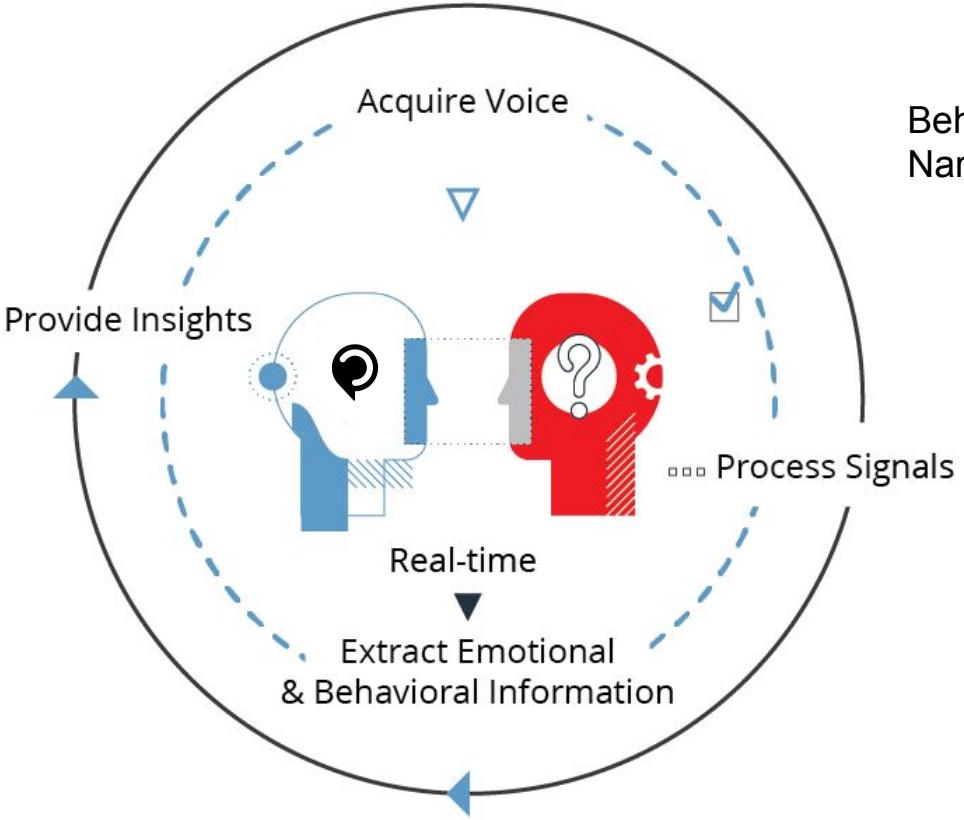
https://medium.com/behavioral-signals-ai/behavioral-signals-what-is-that-367ba0de49d2

BEHAVIORAL SIGNALS

The big picture

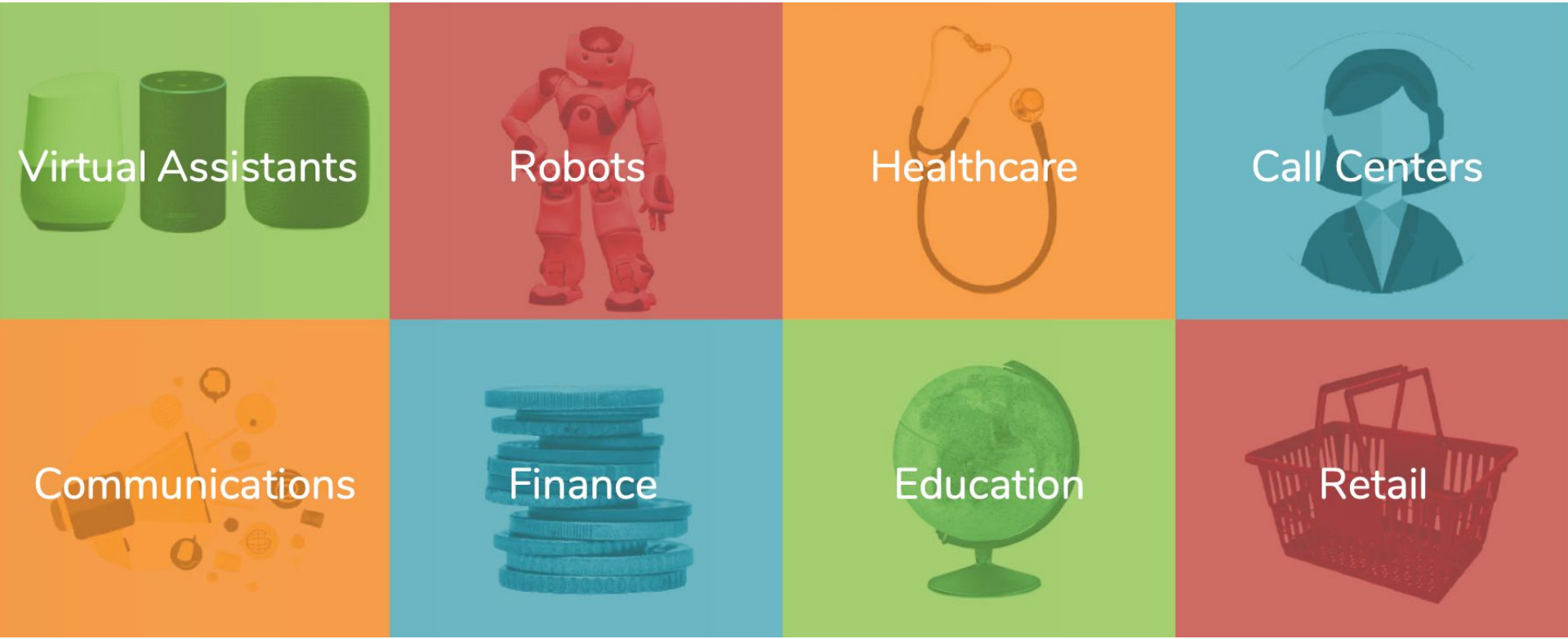


A small slice of the big picture



Behavioral Signal Processing,
Narayanan et al 2009

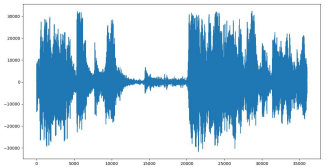
Applications

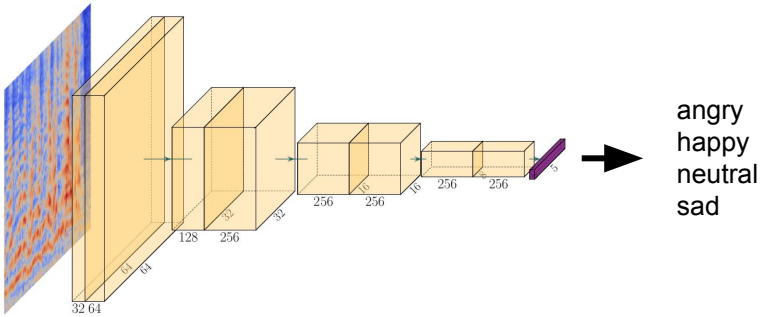


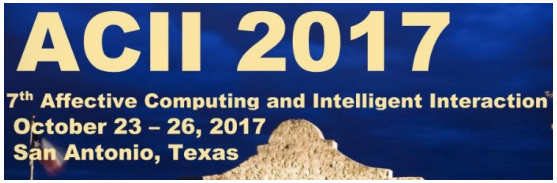
Part I:
Baseline System for Speech Emotion Recognition and Behavioral Tracking: Features and ML Architectures

Speech Emotion Recognition - Methods

- **Goal**: Map audio signals to emotions
- **Features** (Ayadi et al. 2011)
  - LLDs (e.g. MFCCs, Energy, Pitch)
  - **Spectrograms**
- **Classification** (Trigeorgis et al. 2016)
  - Typical supervised models (SVMs)
  - Neural Networks (**CNNs**)
  - Temporal modeling (LSTMs)
- Unlabeled data







Segment-Based Speech Emotion Recognition Using Recurrent Neural Networks

Efthimios Tzinis [1,2], Alexandros Potamianos [1,2]

[1]National Technical University of Athens

[2] Behavioral Signals Technologies

BEHAVIORAL SIGNALS

Sets of Acoustic Features

Local Features: Low level Descriptors (LLDs) over frames, e.g., energy, pitch

(RMS) Root Mean Square, (ZCR) Zero Crossing Rate, (HNR) Harmonics to Noise Ratio, (DDP) Difference of Difference of Periods, (LSP) Line Spectral Pairs, (SHS) Sub-Harmonic Sum, (ACF) Autocorrelation Function, (MFB) Mel Frequency Band.

LLDs	1st Delta	Local Features	Global-Features Applied Functional Sets*
RMS Energy	✓	✓	✗
Quality of Voice	✓	✓	✗
ZCR	✓	✓	✗
Jitter Local	✗	✓	A
Jitter DDP	✗	✓	A
Shimmer Local	✗	✓	A
F0 by SHS	✓	✓	A,C
Loudness	✓	✓	A,B
Probability of Voicing	✓	✓	A,B
HNR by ACF	✓	✓	A,B
MFCCs[0-14]	✓	✓	A,B
LSP Frequency [0-7]	✓	✗	A,B
log MFB [0-7]	✓	✗	A,B
F0 Envelope	✓	✗	A,B

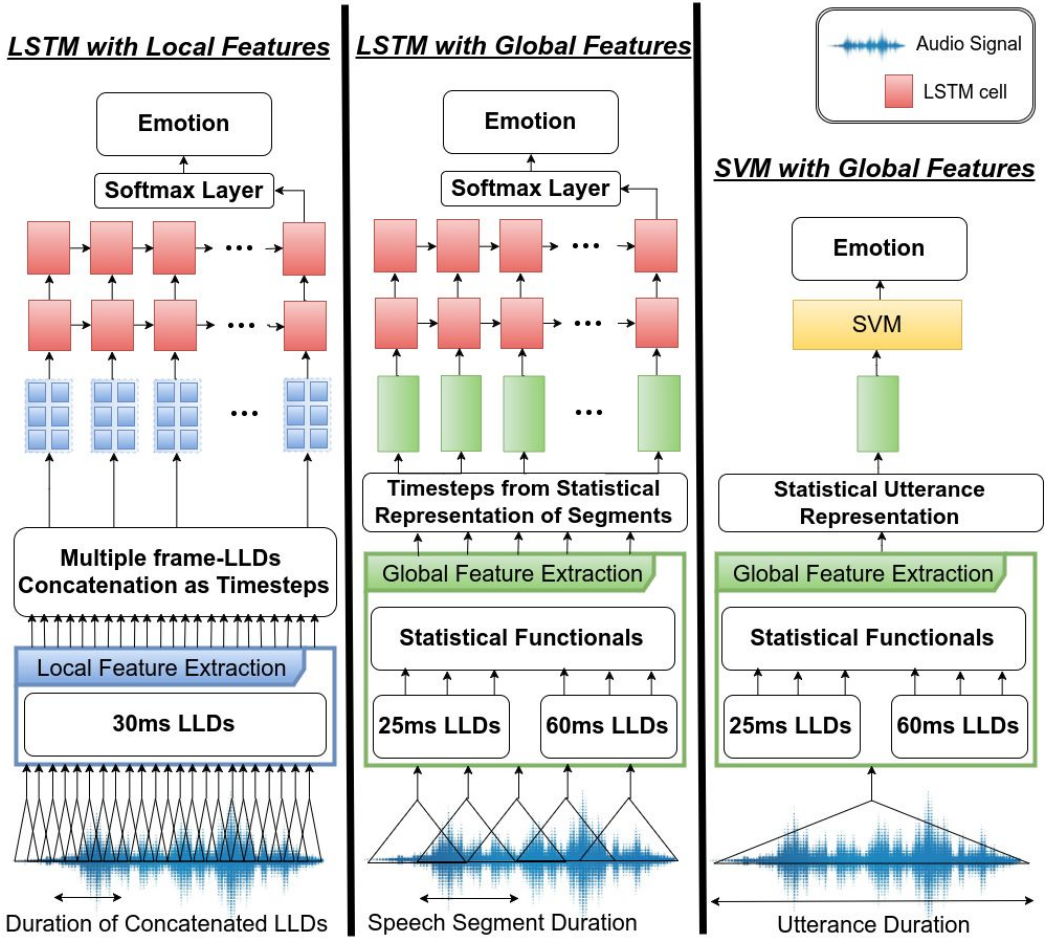
Global Features: statistical functionals over LLDs, e.g., mean, skewness

Statistical Functions	Set
position max/min arithmetic mean, standard deviation skewness, kurtosis linear regression coefficient 1/2 Quadratic & Absolute linear regression error quartile 1/2/3 quartile range 2-1/3-2/3-1 percentile 99 up-level time 75/90	A
percentile 1, percentile range 1-99	B
OnSets Number, Duration	C

BEHAVIORAL SIGNALS

ML models

- SER classification models
- Long Short Time Memory (LSTM) unit
  - Trained with different timesteps (frame or segments features)
- Support Vector Machine (SVM)



Experimental Setup

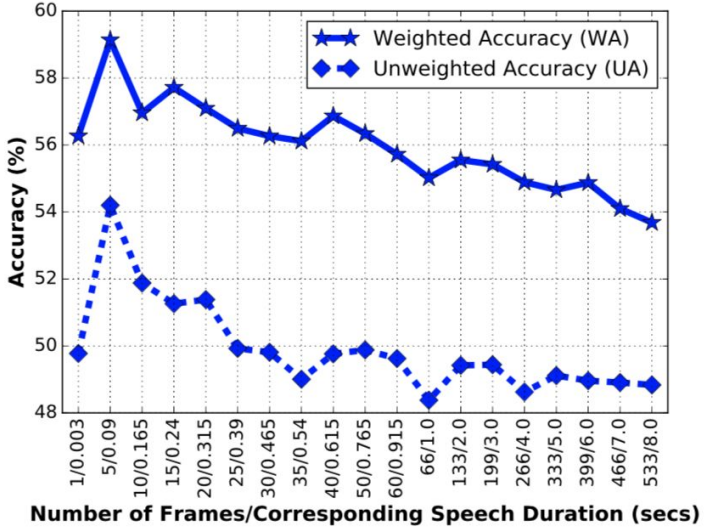
- Database: IEMOCAP
  - 5 Sessions: 2 speakers per session (1 Male, 1 Female)
  - 4 Emotional categories: *Angry*, *Sad*, *Happy*, *Neutral*
- Evaluation Schema:
  - Leave One Session Out (LOSO): 5 folds (4 train, 1 test)
  - Test: 1 speaker for validation and the other for testing
  - Repeat in reverse and compute the average
- Evaluation Metrics:
  - Weighted Accuracy (WA): Percentage of correct classification decisions
  - Unweighted Accuracy (UA): Average of accuracies of all emotional classes

- LSTM Training Setup:
  - 2 Layers: 512 and 256 respectively
  - Nadam optimizer
  - Non-feedback connections' dropout ratio=0.5
  - Local Features: Consecutive frame-vectors (47 LLDs each) in chunks of different lengths corresponding to speech durations ranging from 30ms to 8s
  - Global Features: Statistical representations (1582 features each) segments of lengths ranging from 0.5s to 8s with overlap ratio of 0.5
- SVM Training Setup:
  - Radial Base Function (RBF) kernel
  - Using validation speaker for setting cost coefficient
  - Global features over the whole utterance (1582 features per utterance)

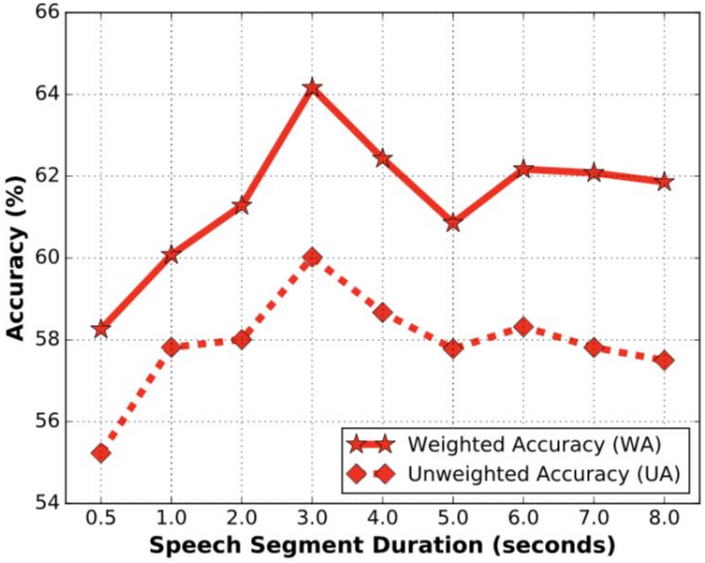
BEHAVIORAL SIGNALS

Results

LSTM with Local Features



LSTM with Global Features



BEHAVIORAL SIGNALS

Comparison with the state-of-the-art

Model	Type of Features	WA (%)	UA(%)
Best LSTM [23]	Spectrogram	61.71	58.05
BLSTM-SUA [8]	LLDs	59.33	49.96
BLSTM-WPA [18]	LLDs	63.5	58.8
BLSTM-ELM [16]	LLDs chunks of 250ms	62.85	63.89

Model	Type of Features	WA (%)	UA(%)
SVM	Statistical over the whole utterance	53.54	49.23
LSTM	LLDs chunks of 90ms	59.14	54.2
LSTM	Statistical over 3 seconds segments	64.16	60.02

[8] Huang, C., W., Narayanan, S., "Attention Assisted Discovery of SubUtterance Structure in Speech Emotion Recognition," INTERSPEECH, pp. 1387–1391, 2016.
[16] Lee, J. and Tashev, I., "High-level feature representation using recurrent neural network for speech emotion recognition," INTERSPEECH, pp. 1537–1540, 2015.
[18] Mirsamadi, S., Barsoum, E. and Zhang, C., (in press), "Automatic Speech Emotion Recognition Using Recurrent Neural Networks With Local Attention," ICASSP, 2017.
[23] Fayek, H., M., Lech, M. and Cavedon, L., (in press), "Evaluating deep learning architectures for Speech Emotion Recognition," Neural Networks, 2017.

BEHAVIORAL SIGNALS

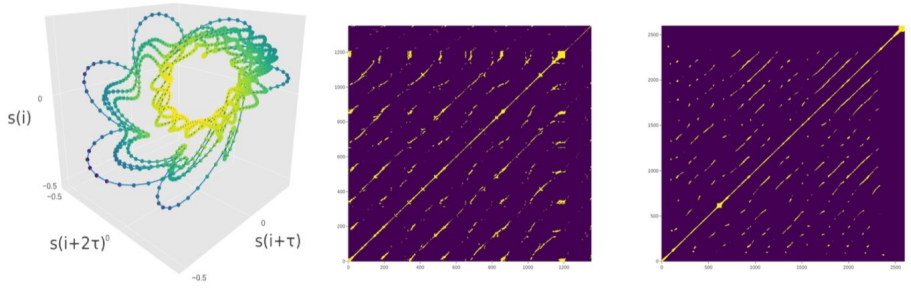
Part II
Feature Extraction and Dimensionality Reduction

Integrating Recurrence Dynamics for Speech Emotion Recognition

Efthimios Tzinis, Georgios Paraskevopoulos, Christos Baziotis,
Alexandros Potamianos

BEHAVIORAL SIGNALS

Phase Space Trajectories and Features



(a) (b) (c)

Figure 1: *(a) Reconstructed PS ($m = 3$, $\tau = 7$) and (b) RP ($\epsilon = 0.15$, Manhattan norm) of 30ms frame corresponding to vowel /e/. (c) RP of Lorenz96 system displaying chaotic behavior [29]*

$$\mathbf{x}(i) = [s(i), s(i+\tau), ..., s(i+(m-1)\tau)]$$

$$\mathbf{R}_{i,j}(\epsilon, q) = \Theta(\epsilon - ||\mathbf{x}(i) - \mathbf{x}(j)||_q)$$

Name	Formulation
Recurrence Rate	$\frac{1}{N^2}\sum_{i,j=1}^{N}\mathbf{R}_{i,j}$
Determinism	$\frac{\sum_{l=d_m}^{N} lP_d(l)}{\sum_{l=1}^{N} lP_d(l)}$
Max Diagonal Length	$max(\{l_i\}_{i=1}^{N_d})$
Average Diagonal Length	$\frac{\sum_{l=d_m}^{N} lP_d(l)}{\sum_{l=d_m}^{N} P_d(l)}$
Diagonal Entropy	$\sum_{l=d_m}^{N}\frac{P_d(l)}{N_d} ln(\frac{N_d}{P_d(l)})$
Laminarity	$\frac{\sum_{l=v_m}^{N} lP_v(l)}{\sum_{l=1}^{N} lP_v(l)}$
Max Vertical Length	$max(\{v_i\}_{i=1}^{N_v})$
Trapping Time	$\frac{\sum_{l=v_m}^{N} lP_v(l)}{\sum_{l=v_m}^{N} P_v(l)}$
Vertical Entropy	$\sum_{l=v_m}^{N}\frac{P_v(l)}{N_v} ln(\frac{N_v}{P_v(l)})$
Max White Vertical Length	$max(\{w_i\}_{i=1}^{N_w})$
Average White Vertical Length	$\frac{\sum_{l=w_m}^{N} lP_w(l)}{\sum_{l=w_m}^{N} P_w(l)}$
White Vertical Entropy	$\sum_{l=w_m}^{N}\frac{P_w(l)}{N_w} ln(\frac{N_w}{P_w(l)})$

Results

Speaker Dependent Results

Features	Model	SAVEE		Emo-DB	
		WA	UA	WA	UA
IS10	SVM	77.1	74.5	88.4	87.2
	LR	74.4	71.8	87.4	86.3
RQA	SVM	66.0	63.0	81.8	80.4
	LR	64.4	61.1	81.9	79.9
RQA+IS10	SVM	77.3	75.5	90.1	88.9
	LR	**80.2**	**77.9**	**93.3**	**92.9**
[14] Spectrogram	SAE	75.4	-	88.3	-
[36] LLDs Stats	ESR	76.3	73.4	88.7	87.9

Speaker Independent Results

Features	Model	SAVEE		Emo-DB	
		WA	UA	WA	UA
IS10	SVM	47.5	45.6	79.7	74.3
	LR	48.5	43.1	76.1	71.9
RQA	SVM	45.6	41.1	70.9	64.2
	LR	47.7	42.3	71.1	67.1
RQA+IS10	SVM	52.5	50.6	82.1	76.9
	LR	**54.0**	**53.8**	80.1	77.5
[36] LLDs Stats	ESR	51.5	49.3	**82.4**	**78.7**
[37] WSFHM+IS10	SVM	50.0	-	81.7	-

BEHAVIORAL SIGNALS

GRAZ – AUSTRIA
SEPTEMBER 15th – 19th 2019

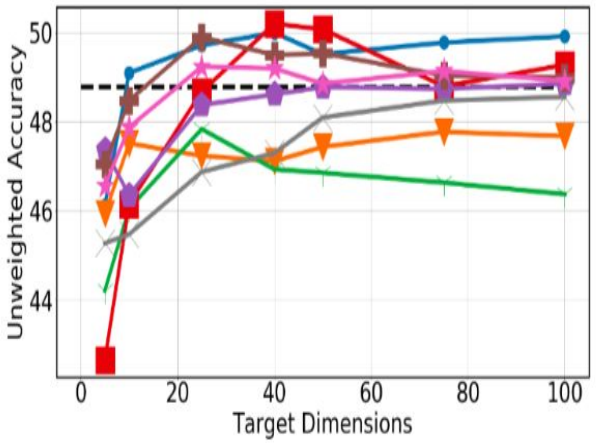
INTERSPEECH
2019

Unsupervised Low-Rank Representations for Speech Emotion Recognition

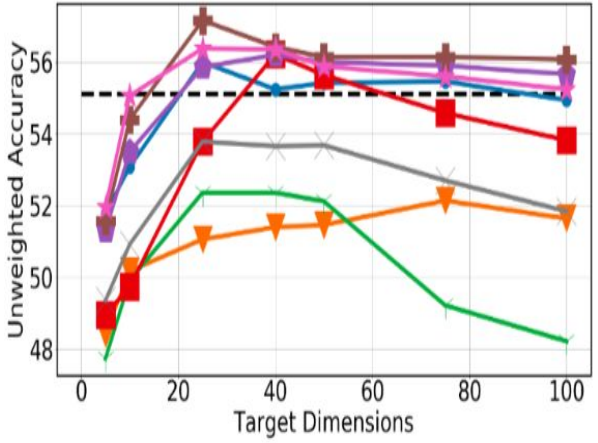
Georgios Paraskevopoulos, Efthymios Tzinis, Nikolaos Ellinas, Theodoros Giannakopoulos, Alexandros Potamianos

BEHAVIORAL
SIGNALS

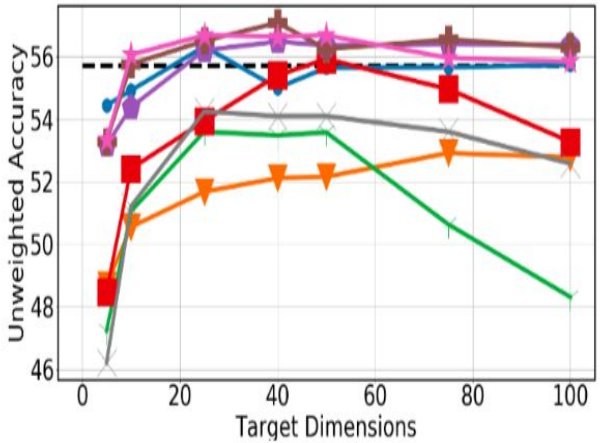
Performance of Dimensionality Reduction Algorithms (k-NN)



(d) RQA feature set on IEMOCAP



(e) IS10 feature set on IEMOCAP



(f) Fused feature set on IEMOCAP

BEHAVIORAL SIGNALS

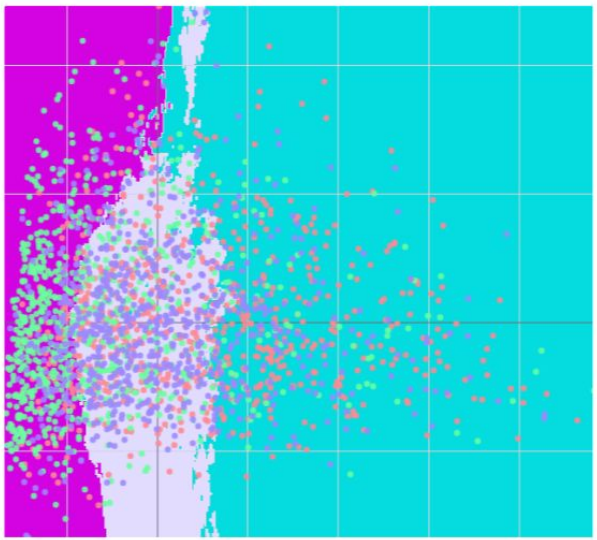
Performance of Dimensionality Reduction on IS10 dataset

- From 1582 down to 25 dimensions

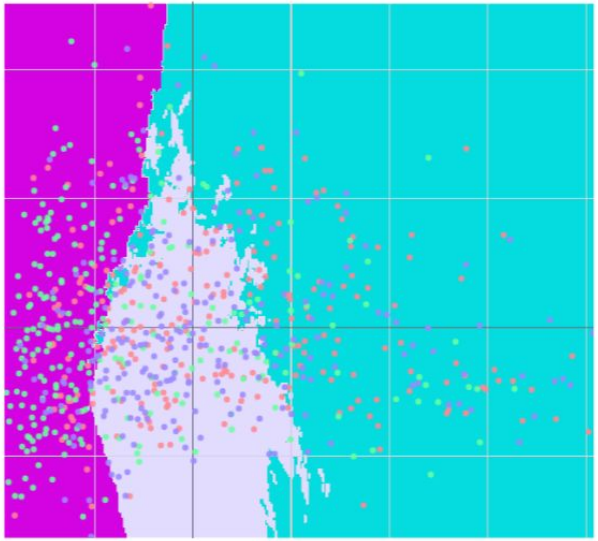
	SVM (linear)	SVM (rbf)	kNN	LR
Pattern S. MDS	**56.0**	57.5	56.5	55.4
SMACOF MDS	55.8	**58.5**	**56.7**	**55.8**
PCA	55.8	57.7	56.2	**55.8**
ISOMAP	52.3	52.5	51.7	52.2
LLE	53.4	54.2	53.6	53.2
Modified LLE	54.6	47.0	53.9	55.5
Spectral Emb.	54.1	54.3	54.2	55.1
Autoencoder	55.4	57.8	56.3	55.5
Original 1582D	54.7	**59.8**	55.7	**56.9**

BEHAVIORAL SIGNALS

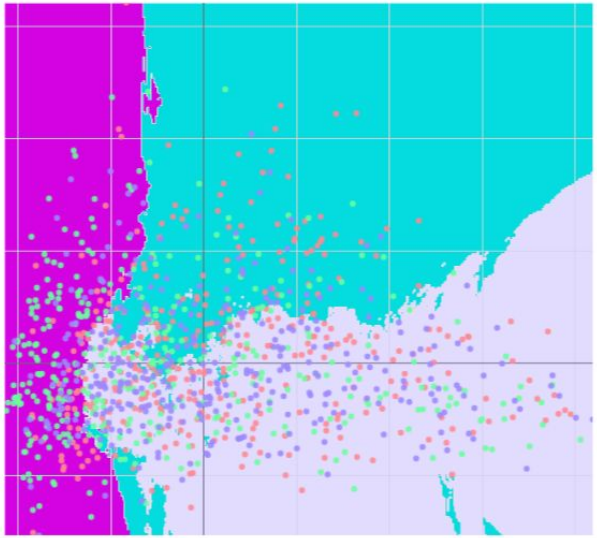
Cross-domain Decision Boundaries



(a) Series



(b) Movies



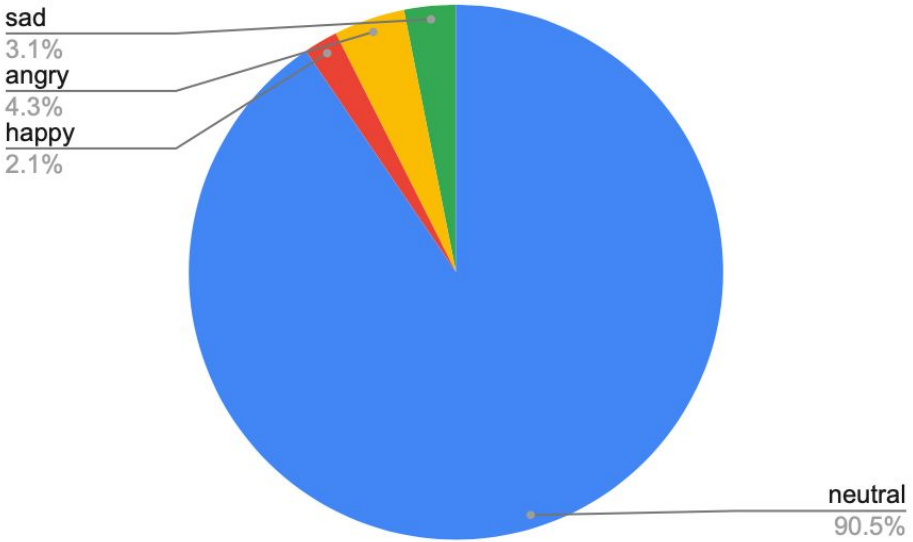
(c) Interviews

anger (blue dots) vs happy (green dots) vs sad (red dots)
with k-NN decision boundaries

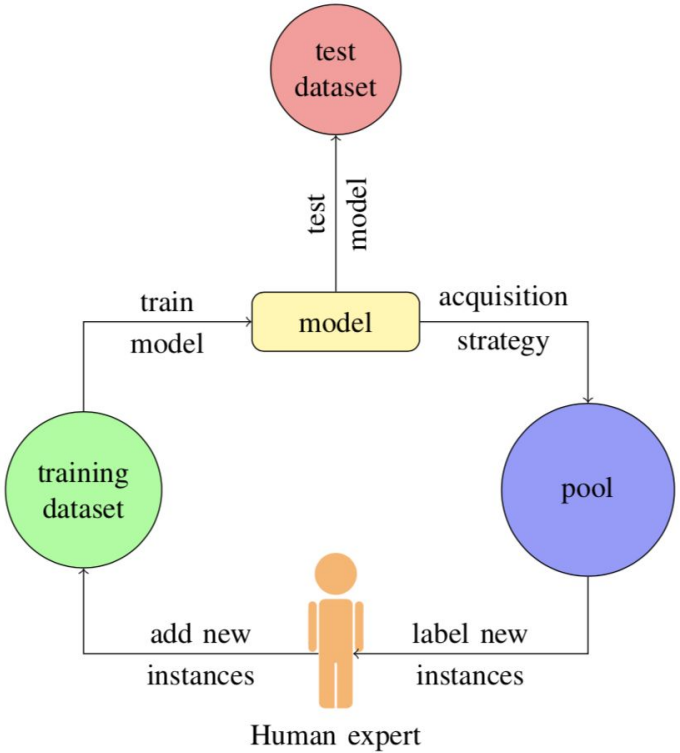
Part III
Data Sparsity: Data Selection and Data Augmentation

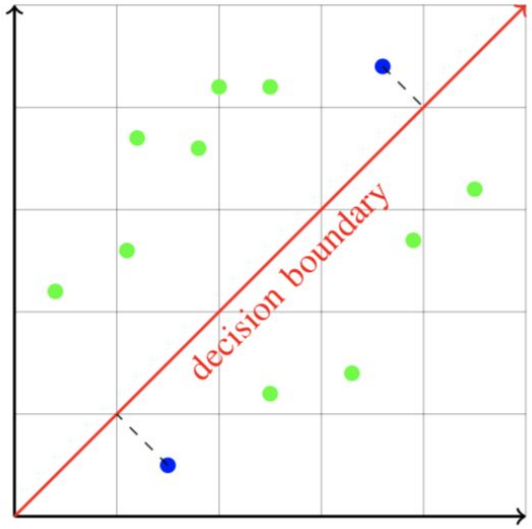
Problem: Data imbalance in speech emotion recognition

- Data in the wild are highly imbalanced
  - **Sparse** non-neutral samples
- Annotation is hard
  - Bias towards neutral
  - **Low** inter-annotator agreement



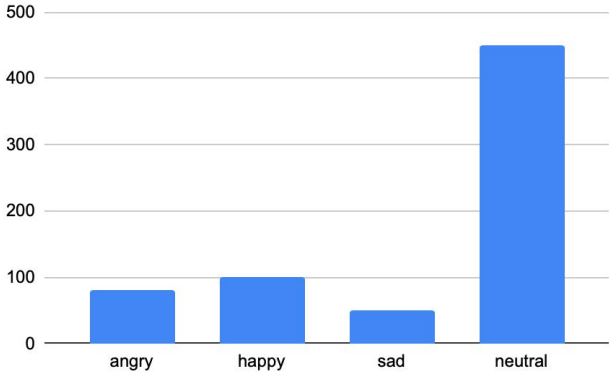
Possible Solutions: Active Learning

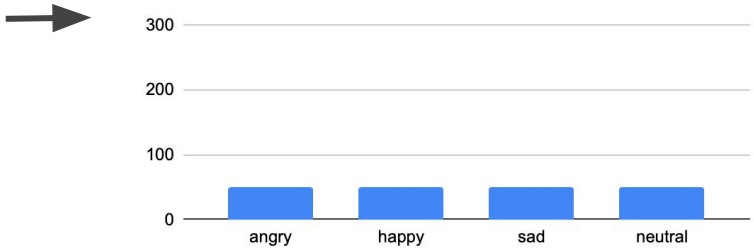




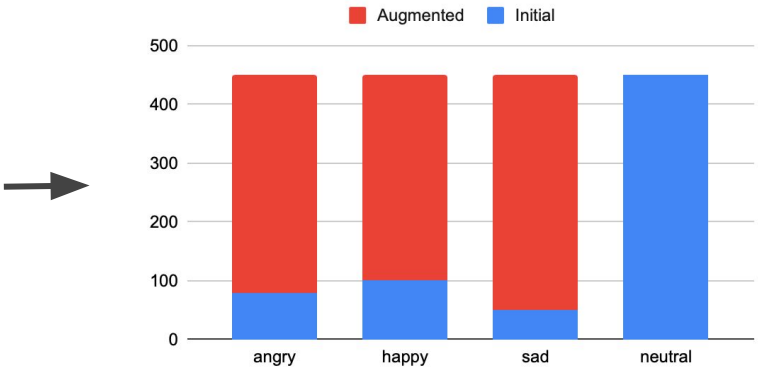
Informative samples lying closest to the decision boundary.

Possible Solutions: Subsampling Vs Augmentation





Subsampling



Augmentation

GRAZ – AUSTRIA
SEPTEMBER 15th – 19th 2019

INTERSPEECH
2019

Data Augmentation using GANs for Speech Emotion Recognition

Aggelina Chatziagapi, Georgios Paraskevopoulos, Dimitris Sgouropoulos, Georgios Pantazopoulos, Malvina Nikandrou, Theodoros Giannakopoulos, Athanasios Katsamanis, Alexandros Potamianos, Shrikanth Narayanan

BEHAVIORAL SIGNALS

Data Augmentation

- Generate samples for underrepresented emotion classes → How?
- **Signal-based** (Schlüter et al. 2015, Salamon et al. 2017, Etienne et al. 2018)
  - Add **noise**
    - Gaussian
    - Ambient sounds (“real noise”) *
  - **Transformations**
    - Pitch shift
    - Time stretch (Aldeneh et al. 2017)
- **Model-based**
  - Use (generative) learning approaches
    - Balancing GAN (Mariani et al. 2018)

* From ESC-50 and FMA datasets, including audio events and music sounds

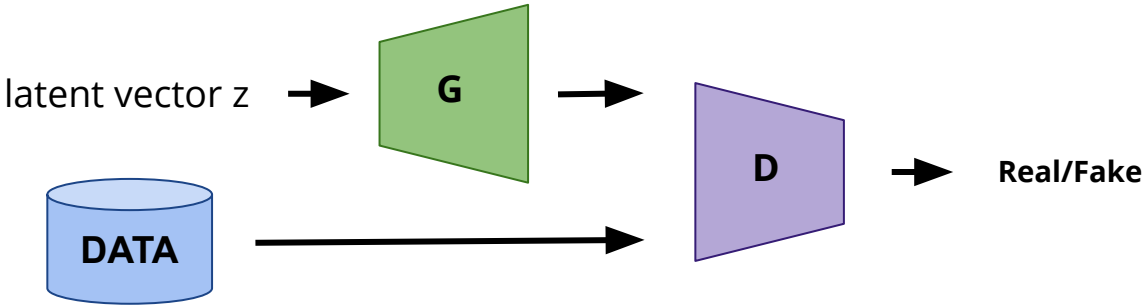
Generative Adversarial Networks (GANs)

- **Goal: Sample generation**

Real (CIFAR-10)

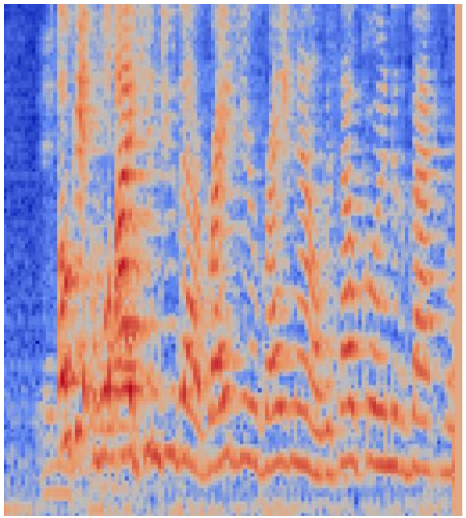
Generated
(Salimans et al. 2016)

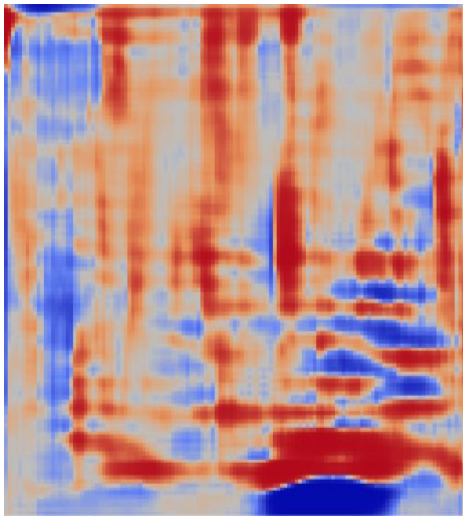
- **2 competing networks - Minimax game** (Goodfellow et al. 2014)



- Discriminator *D* maximizes the probability of assigning correct labels (real vs fake)
- Generator *G* maximizes the probability of *D* making a mistake

Spectrogram Generation: BAGAN

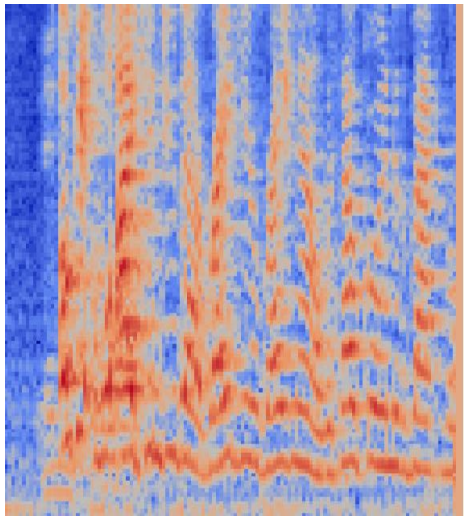




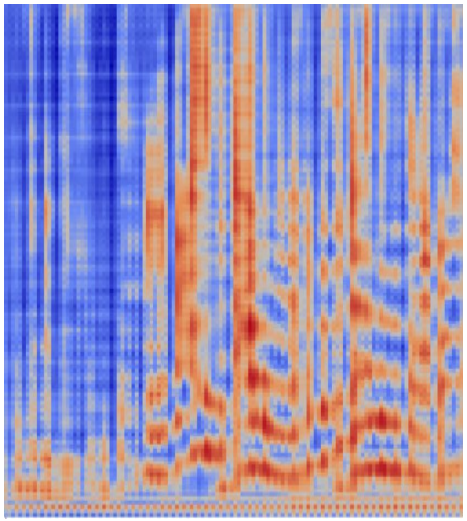
Real

BAGAN
(https://github.com/IBM/BAGAN)

Spectrogram Generation: Proposed



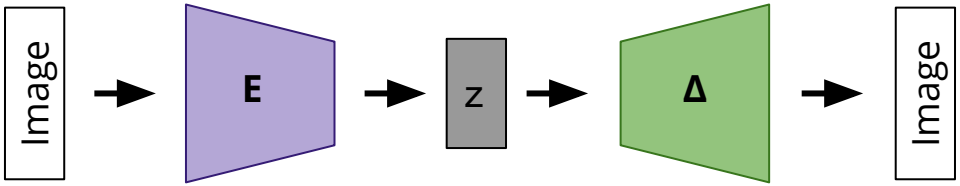
Real



Proposed

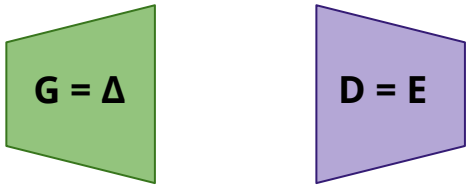
Balancing GAN (BAGAN) - Autoencoder Training

- Autoencoder training (unsupervised)
- Learn weights close to a good solution
  - Faster convergence
  - Avoid mode collapse



Balancing GAN (BAGAN) - GAN Initialization

- Transfer weights to GAN
  - Replace the last dense layer of *E* with dense layer and softmax for *D*

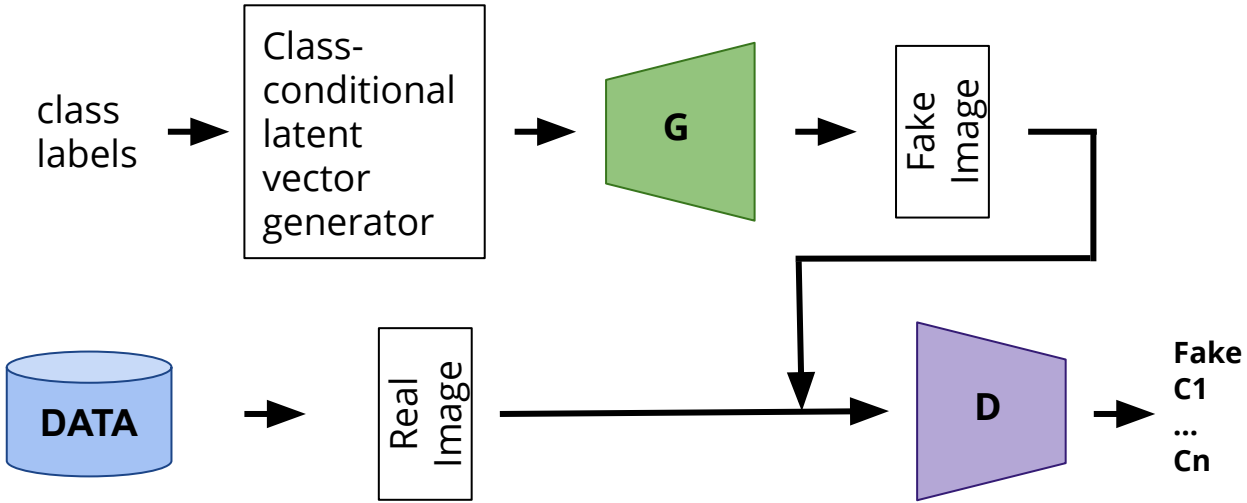


- Class conditioning
  - Per class c: Compute mean (μ_c) and covariance matrix (Σ_c) of latent vector z
  - Model each class with a multivariate normal distribution $N_c = N(\mu_c, \Sigma_c)$

⇒ Class-conditional latent vector generator

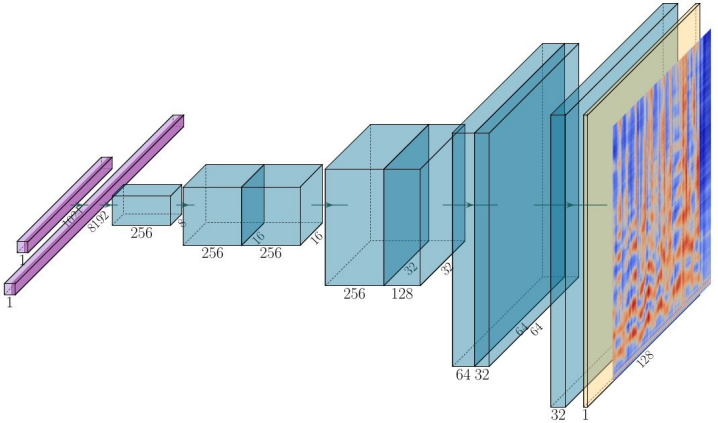
Balancing GAN (BAGAN) - GAN Fine-tuning

- Train GAN using both majority and minority classes
  - D: match real images with the correct class labels and the generated with the fake label
  - G: match the labels generated by D with the labels used to generate images

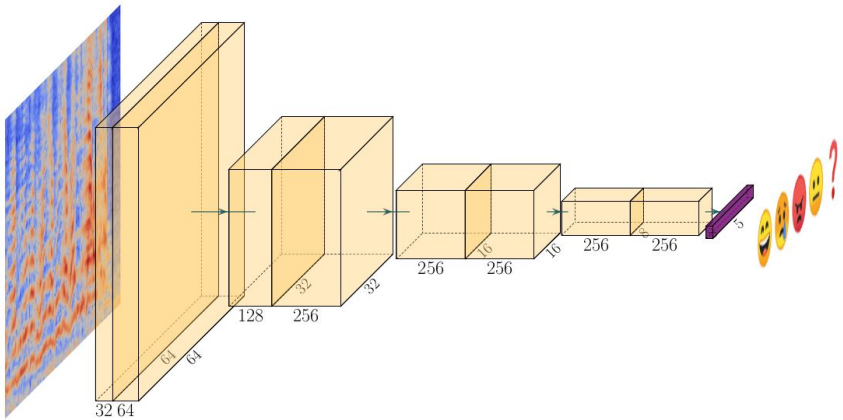


- Use *G* to generate spectrograms for each minority class

Proposed Method - Architecture



Generator (G)



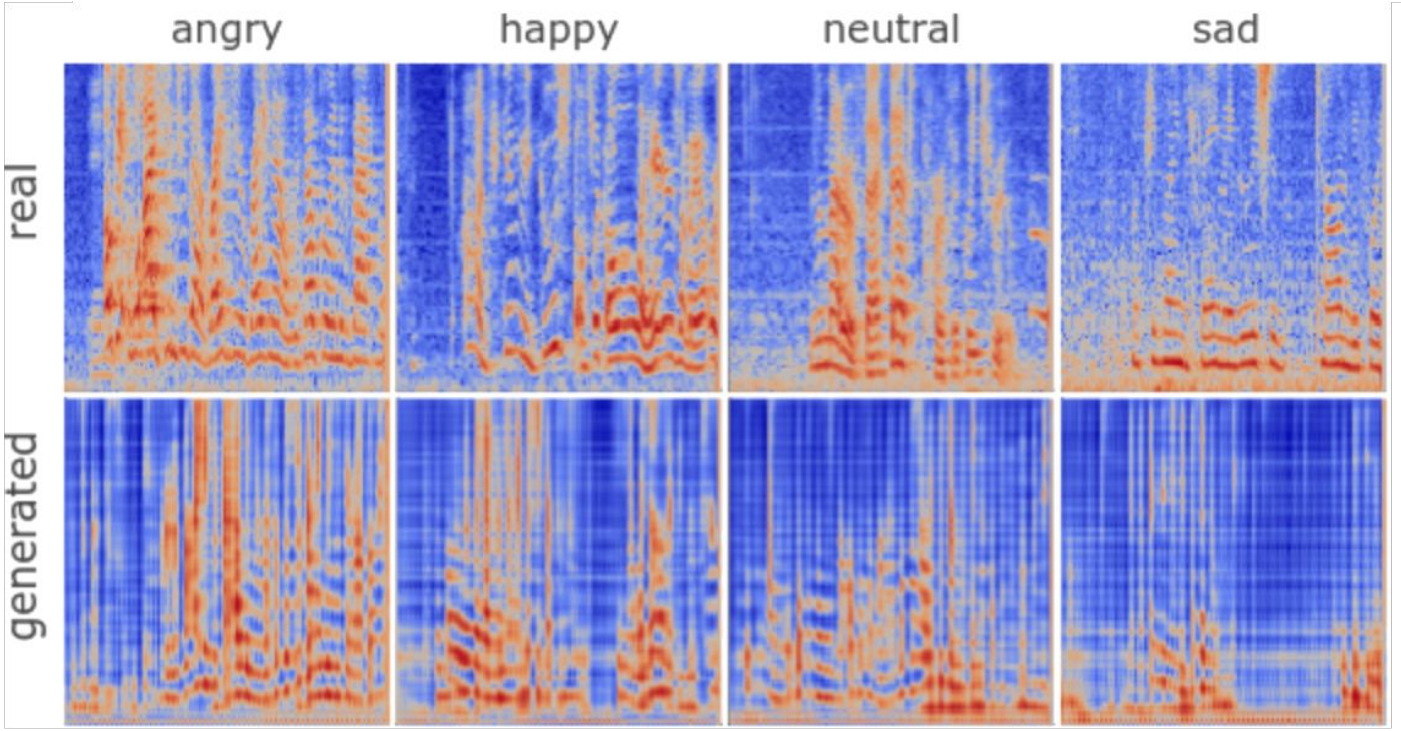
Discriminator (D)

Contributions

The proposed GAN-based augmentation method is differentiated from BAGAN:

- Use **Transposed Convolutional layers** instead of Upsampling (*G*)
- **Leaky ReLU** for all intermediate activation layers
- **Batch Normalization** and **Dropout** for regularization
- Feed *D* with **separate batches of real and fake** images

Real and Generated spectrograms



Real and generated spectrograms for each emotion class, from IEMOCAP dataset.

Experimental Setup - Datasets

- Datasets:
  - **IEMOCAP**
    - classes: angry, happy, neutral, sad
    - 5531 utterances (7 hours)
  - **FEEL-25k**
    - classes: ambiguous, angry, happy, neutral, sad
    - 25k utterances (49 hours)
    - diverse domains
- Imbalance strategy:
  - **IEMOCAP:** Simulate imbalance issue - Remove 80% of the target emotional class from training set
  - **FEEL-25k:** Already imbalanced (sad/neutral = ⅕)

Experimental Setup

- Train-test split:
  - **IEMOCAP:** 5 fold cross-validation (leave-one-session-out)
  - **FEEL-25k:** 80%-20% shuffle split. Test on separate test set (50k utterances - 100 hours)
- Feature Extraction:
  - Mel-spectrograms
  - Per 3-sec segment: 128x128 image
  - Min-max normalization [-1, 1]
- Apply augmentation approaches on training set
- Classification:
  - VGG19 (Simonyan et al. 2015)
  - Majority voting

Results - IEMOCAP

Method	Angry	Happy	Sad	Average
Imbalanced	47.8	52.2	46.9	49.0
Signal-based	49.7	49.6	47.6	49.0
Proposed approach	**53.5**	**55.2**	**52.1**	**53.6**

IEMOCAP performance (UAR %)

Each column: Simulation (remove 80% of emotion class samples in training set and then augment it)
This results in a tiny amount of samples (~180) for the minority class.

Results - FEEL-25k

Method	UAR	F-score
Initial	52.3	52.7
Subsampling	49.4	48.9
Signal-based	51.2	50.0
Proposed approach	**54.6**	**55.0**

FEEL-25k performance (UAR & F-score %)

Conclusions

- Address the imbalance issue in real-world SER using GANs
- Relative performance improvement: **5%** (FEEL-25k) to **10%** (IEMOCAP)
- Signal-based augmentation did not work
- Future work:
  - Temporal modeling: combine CNN with LSTM
  - Improve spectrogram quality (stripes)
  - Compare with data-driven baselines (e.g. VAE) and transfer learning

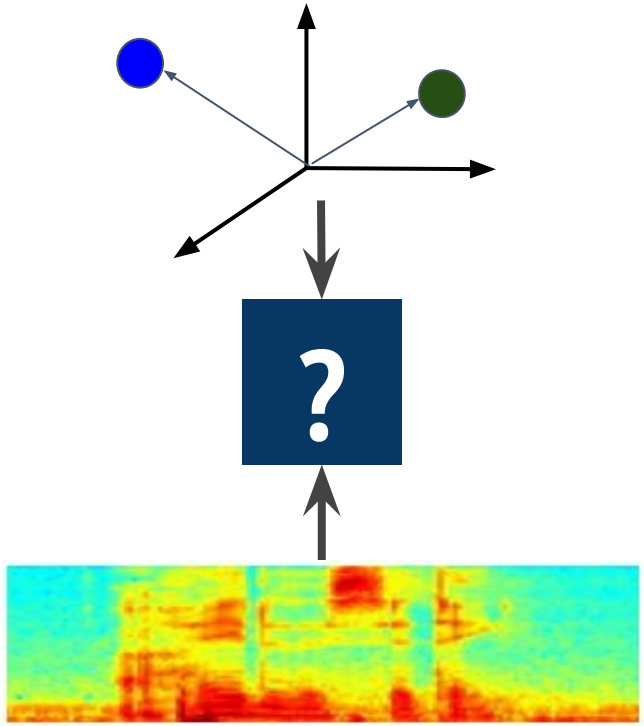
Part IV
Multimodal Fusion

Why go multimodal ?

- Human interactions:
  - Naturally involve multiple modalities
- Vocal communication:
  - Information about the emotional state of the speaker is evident in both text and audio
- Textual Modality:
  - **What is said**
- Acoustic Modality:
  - **How things are being said**
- Human-Machine Interaction Systems:
  - Should encapsulate both the semantic and affective information of a message

Challenges that arise

- We need to take into account cross-modal interactions:
  - Time-dependent
  - Time-independent
- Modalities operate at different timescales:
  - Text: character-word-sentence level
  - Audio: segment-utterance level
- Heterogenous representations
  - Text: word embeddings (GloVe, ...)
  - Audio: low level features (MFCCs, ...)
- **Fusion Strategy**
  - **What** makes sense to be fused
  - **How** it can be fused



GRAZ – AUSTRIA
SEPTEMBER 15th – 19th 2019

INTERSPEECH
2019

Deep Hierarchical Fusion with application in Sentiment Analysis

Efthymios Georgiou [1,2], Charilaos Papaioannou [1], Alexandros Potamianos [1,2]

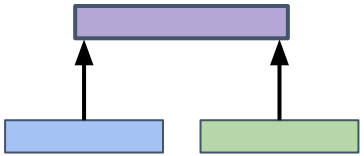
[1]National Technical University of Athens

[2] Behavioral Signals Technologies

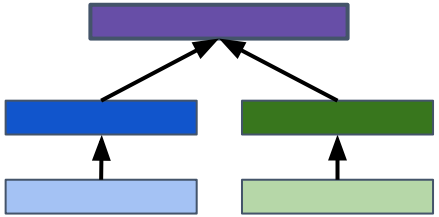
BEHAVIORAL SIGNALS

How to go multimodal ?

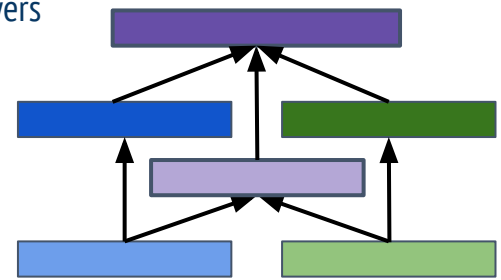
- **Goal:** Join information from two modalities to perform a prediction
  - This is essentially **fusion**
- Fusion Strategies:
  - Early: combines feature level representations to a unified vector
  - Late: separately trained classifiers fuse their decisions
  - Hybrid: exploits both early and late fusion strategies
- Dense Fusion (Hu et al. 2019)
  - Combines representations in different (early and intermediate) shared layers



Early Fusion



Intermediate Fusion

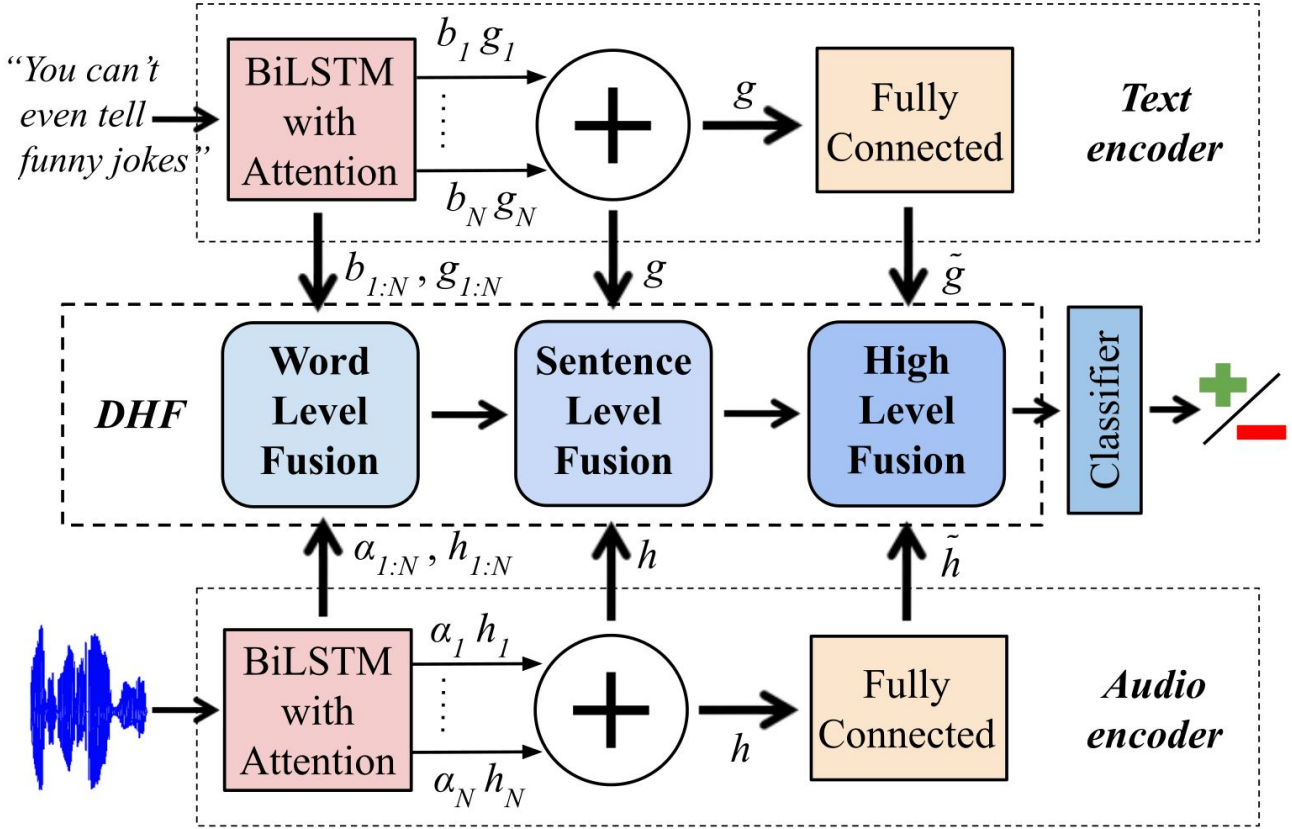


Dense Fusion

Solutions

- Synchronize audio with text modality
  - Word-level alignment:
    - Text: naturally operates on the word-level
    - Audio: how the word is uttered
- **Deep Hierarchical Fusion**
  - Motivation: Deep Learning and Human Cortical networks build upon hierarchies of concepts and representations
  - Key ideas:
    - **Fuse at multiple levels** to better capture cross-modal correlations
    - **Re-use fused** multimodal representations
    - **Re-use unimodal** representations

DHF


"You can't even tell funny jokes"
BiLSTM with Attention
$b_1 g_1$
$b_N g_N$
g
Fully Connected
Text encoder
$b_{1:N}, g_{1:N}$
g
$\tilde{g}$
DHF
Word Level Fusion
Sentence Level Fusion
High Level Fusion
Classifier
$\alpha_{1:N}, h_{1:N}$
h
$\tilde{h}$
BiLSTM with Attention
$\alpha_1 h_1$
$\alpha_N h_N$
h
Fully Connected
Audio encoder

DHF Architecture

- The proposed architecture consists of three parts
  - Text Encoder: BiLSTM with attention mechanism followed by dense layers
  - Audio Encoder: Same as Text Encoder
  - DHF: The fusion network
- Two graphical directions of the information flow
  - Vertical: different unimodal representations that are fed to DHF
  - Horizontal: the constant forward propagation of multimodal information
- DHF Modules:
  - **Word-Level-Fusion**:
    - Goal: Capture time-dependent cross-modal interactions
    - Architecture: BiLSTM with attention mechanism
    - Fusion Rule: *audio hid state* || *text hid state* || *audio hid state* ⊙ *text hid state*
  - **Sentence-Level-Fusion**:
    - Goal: Fuse sentence-level representations from audio, text and word-level stage
    - Architecture: Fully Connected Layers
  - **High-Level-Fusion** (same as Sentence-Level)

BEHAVIORAL SIGNALS

Contributions

- DHF performs fusion at different interconnected layers
  - Introducing multiple learning paths and thus **better capturing cross-modal dependencies**
- The different fusion **layers are hierarchically arranged**
- The fused representations are fed forward and re-combined with unimodal ones
  - Introducing **depth** to the architecture and exploiting the notion of **re-use**
- The proposed architecture is general
  - **Task independent**
  - **Can be extended to arbitrary depth**

BEHAVIORAL SIGNALS

Experimental Setup

- Dataset
  - **CMU-MOSI**
    - Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos
    - 93 videos of movie reviews with **2199 opinion segments**
    - 89 distinct speakers (48 male, 41 female)
    - **Audio**-visual data **with transcriptions** and sentiment labels
- Data pre-processing
  - Word-Level Alignment
    - Get the exact time-stamps of every word
  - Textual Representation
    - GloVe embeddings (300d)
  - Acoustic Features:
    - MFCCs, pitch tracking, voiced/unvoiced segment, …
    - The average acoustic vector is obtained for every word (72d)

Both the data gathering and feature extraction along with the word-level alignment, were performed using the ***mmsdk-framework*** (https://github.com/A2Zadeh/CMU-MultimodalSDK)

BEHAVIORAL SIGNALS

Baseline Methods

- **C-MKL** (Poria et al., 2016)
  - Uses a deep **CNN** structure to capture **high-level features** and fuses them through **multiple kernel learning**, which combines modalities by grouping them and assigning a kernel to each one of the groups.
- **TFN** (Zadeh et al., 2017)
  - Uses **outer product between different modality tensors** to capture unimodal and bimodal interactions. Results on sentiment analysis task.
- **FAF** (Gu et al., 2018)
  - Uses a **hierarchical attention strategy for each modality**. A fine-tuning attention mechanism is used to fuse time-dependent representations and its output is fed to a **CNN which performs the final decision**. Also uses only text and audio for affective analysis.
- **MFM** (Tsai et al., 2019)
  - A **GAN**, which defines a **joint distribution over multimodal data**. It takes into account both the generative and the discriminative aspect and aims to generate missing modality values, while projecting them into a common learned space. The **multimodal discriminative factor achieves state-of-the-art** results.

BEHAVIORAL SIGNALS

Experiment 1: Comparison with State-Of-The-Art

Task	Binary		5 class	7 class
	Acc(%)	F1	Acc(%)	Acc(%)
CMK-L	73.6	75.2	-	-
TFN	75.2	76.0	39.6	-
FAF	76.4	76.8	-	-
MFM	76.4	76.3	-	35.0
DHF	**76.9**	**76.9**	**45.47**	**37.14**

- Sentiment Analysis Performance
- **DHF outperforms current SOTA** by a small margin in the binary task (which is the most well studied)
- Other models' results have been reported in the respective papers
- Relative Improvement: 0.5%

BEHAVIORAL SIGNALS

Experiment 2: Fusion Contribution

Model	FAF	TFN	DHF
	Acc(%)	Acc(%)	Acc(%)
Text	75.0	74.8	73.8
Audio	60.2	65.1	63.3
Fusion	76.4	75.2	76.9
$\Delta Fusion$	↑1.4	↑0.4	↑**3.1**

- Even though unimodal classifiers do not outperform baseline ones, **DHF boosts the performance by a significantly larger margin** compared to other methods
- Relative improvement: 3.1%

BEHAVIORAL SIGNALS

Experiment 3: Ablation Study

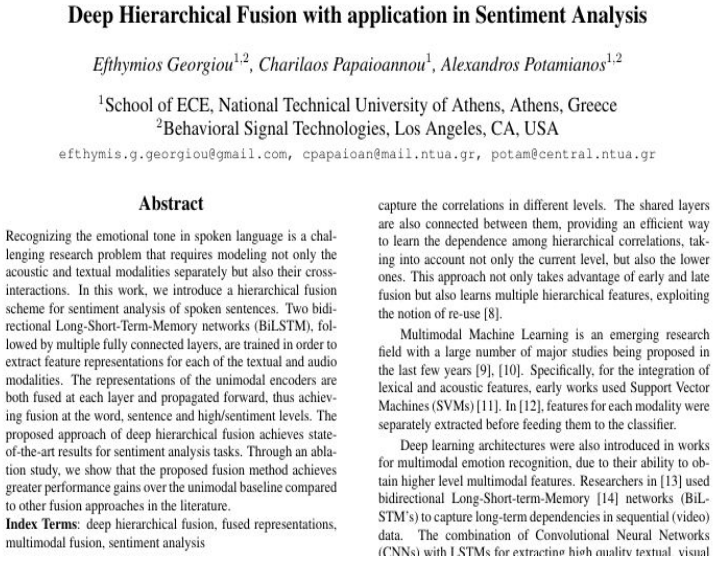
Model	Accuracy(%)	F1
DHF *No High-Level*	75.0	74.8
DHF *No Sent-Level*	75.5	75.4
DHF *No Word-Level*	75.7	75.6
DHF	76.9	76.9

- A level of hierarchy is being subtracted each time to deduce the contribution of the different DHF modules
- **The higher levels of hierarchy are the most important**
- The result denotes that higher-level modules extract more useful representations

BEHAVIORAL SIGNALS

Conclusions

- A **deep hierarchical fusion scheme** is proposed
  - Applied to multimodal (text & audio) sentiment analysis
  - Uses **three levels of hierarchy**.
- DHF achieves:
  - State-Of-The-Art results on the CMU-MOSI database.
  - Relative fusion performance boost (ΔFusion) of **3.1%**
- The proposed **method is general** and can be applied to
  - **Any multimodal task**
  - **Arbitrary depth**



Deep Hierarchical Fusion with application in Sentiment Analysis

Efthymios Georgiou[1,2], *Charilaos Papaioannou*[1], *Alexandros Potamianos*[1,2]

[1]School of ECE, National Technical University of Athens, Athens, Greece
[2]Behavioral Signal Technologies, Los Angeles, CA, USA

efthymis.g.georgiou@gmail.com, cpapaioan@mail.ntua.gr, potam@central.ntua.gr

Abstract

Recognizing the emotional tone in spoken language is a challenging research problem that requires modeling not only the acoustic and textual modalities separately but also their cross-interactions. In this work, we introduce a hierarchical fusion scheme for sentiment analysis of spoken sentences. Two bidirectional Long-Short-Term-Memory networks (BiLSTM), followed by multiple fully connected layers, are trained in order to extract feature representations for each of the textual and audio modalities. The representations of the unimodal encoders are both fused at each layer and propagated forward, thus achieving fusion at the word, sentence and high/sentiment levels. The proposed approach of deep hierarchical fusion achieves state-of-the-art results for sentiment analysis tasks. Through an ablation study, we show that the proposed fusion method achieves greater performance gains over the unimodal baseline compared to other fusion approaches in the literature.

Index Terms: deep hierarchical fusion, fused representations, multimodal fusion, sentiment analysis

capture the correlations in different levels. The shared layers are also connected between them, providing an efficient way to learn the dependence among hierarchical correlations, taking into account not only the current level, but also the lower ones. This approach not only takes advantage of early and late fusion but also learns multiple hierarchical features, exploiting the notion of re-use [8].

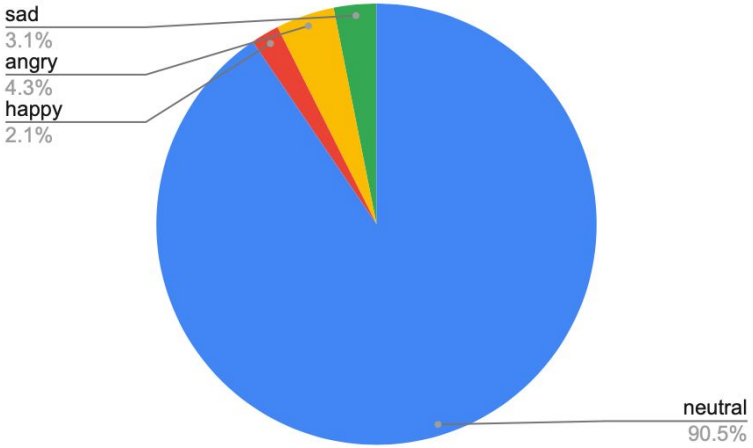
Multimodal Machine Learning is an emerging research field with a large number of major studies being proposed in the last few years [9], [10]. Specifically, for the integration of lexical and acoustic features, early works used Support Vector Machines (SVMs) [11]. In [12], features for each modality were separately extracted before feeding them to the classifier.

Deep learning architectures were also introduced in works for multimodal emotion recognition, due to their ability to obtain higher level multimodal features. Researchers in [13] used bidirectional Long-Short-term-Memory [14] networks (BiLSTM's) to capture long-term dependencies in sequential (video) data. The combination of Convolutional Neural Networks (CNNs) with LSTMs for extracting high quality textual, visual

Part V
Real-World issues: Emotion Recognition in the Wild

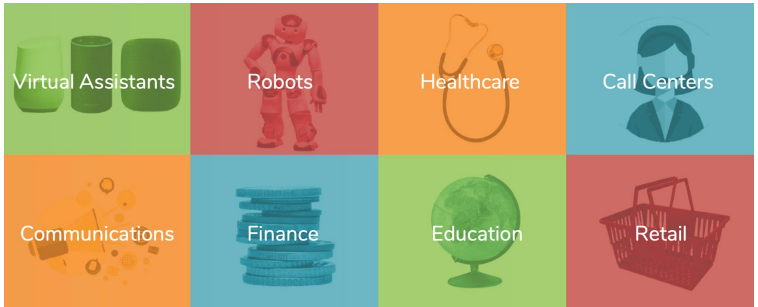
Data Imbalance

- **Distribution** of emotion and behavioral labels varies widely across application domains. Why?
  - Life is boring, movies are exciting, voice assistants are god awful!
  - 5x more emotion in movies than in our daily lives
  - Main emotion present in human-machine interaction currently is frustration
- Challenges towards creating a universal emotion recognizer (across all domains)
  - Tuning the **operation point** across domains (precision/recall trade-off) at the equal error rate
  - F-score performance is lower for domains with rare emotional events



Recording Conditions

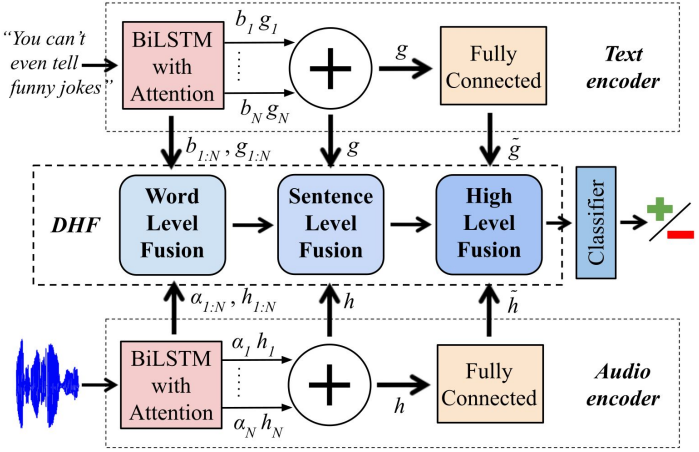
- Recording conditions vary widely across application domains
  - [lab] Production quality data, e.g., movies, TV series, high microphone quality and acted speech
  - [lab] Near production quality data, e.g., interview, vblogs, good microphone quality but spontaneous speech
  - [wild] Average recording quality (far-field but high-quality microphone, background noise), e.g., voice assistants, robotics
  - [wild] Poor and variable recording quality (speakerphone, background noise, bluetooth, cell-phone), e.g., customer-side in call-centers
- ASR performance ranges from 60%-90+% word-accuracy
- SER ranges from 60-80% F-score



BEHAVIORAL SIGNALS

Speech Recognition Errors

- **ASR errors** affect emotion recognition performance from text
  - Roughly a linear relationship between word error rate and F-score for emotion recognition from text
  - Every additional 10% word error rate causes about a 5% drop in emotion recognition F-score
- As a result **audio** is much more reliable than **text** for emotion recognition in the wild
  - Vice-versa in the lab
- Absolute and audio-text fusion performance of multimodal emotion recognition varies per domain



Data Diversity: Duration, Context, Localization

- Utterance duration
  - Longer for human-human vs human-machine interaction
- **Duration** of interaction is typically much longer in the lab than in the wild offering opportunities for
  - Better online adaptation
  - More accurate baselining of the neutral emotional state of a speaker
  - Modeling speaker emotional/behavioral dynamics
- Frequency and strength of emotional expression varies widely in different **cultures**
- Emotion is expressed differently in different **social** (business, personal) and **application** context (retail, healthcare)
- Emotion often needs to be interpreted in the context of **semantics** (what is being said)

Other Considerations

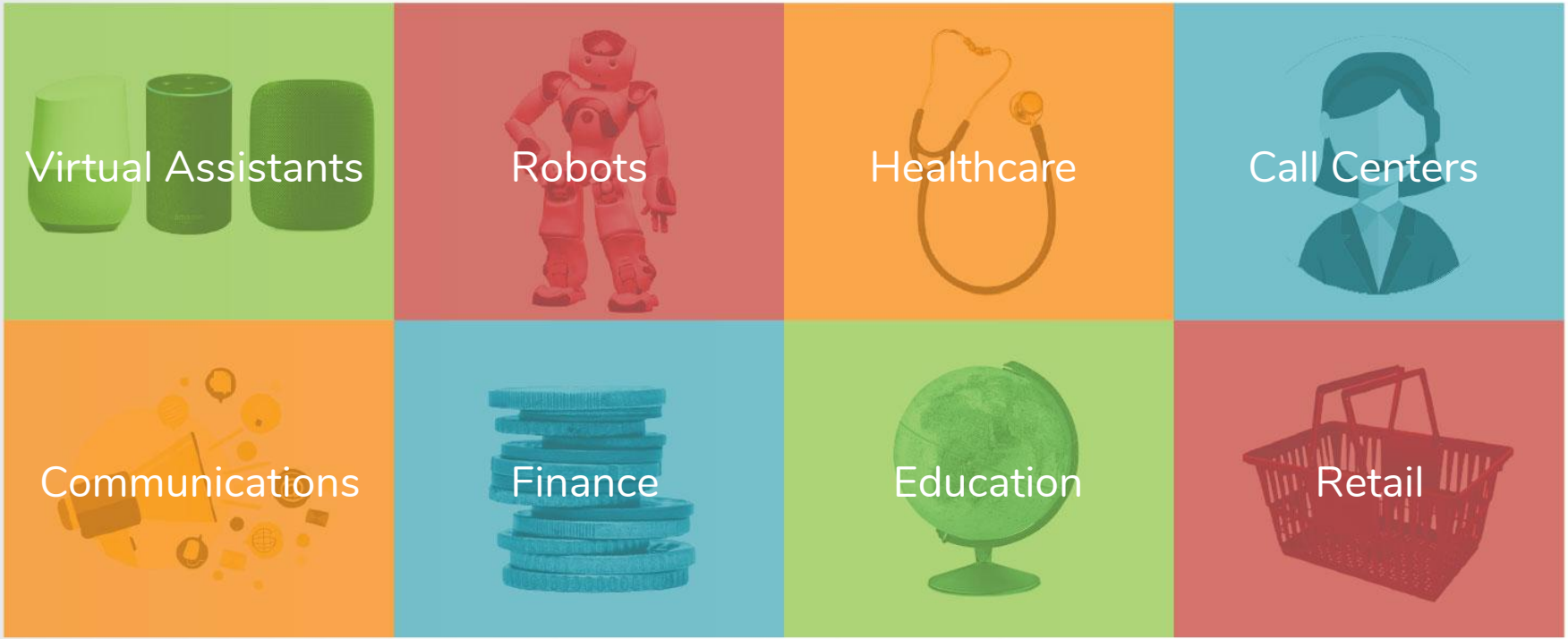
- Speed vs. accuracy
- SVMs vs DNNs
- Privacy / ethical issues

BEHAVIORAL SIGNALS

Part VI
Applications and Demos

Industries

BEHAVIORAL SIGNALS



We Introduce Emotion AI in your applications with

a robust evolving API

+

a team of scientists

BEHAVIORAL SIGNALS

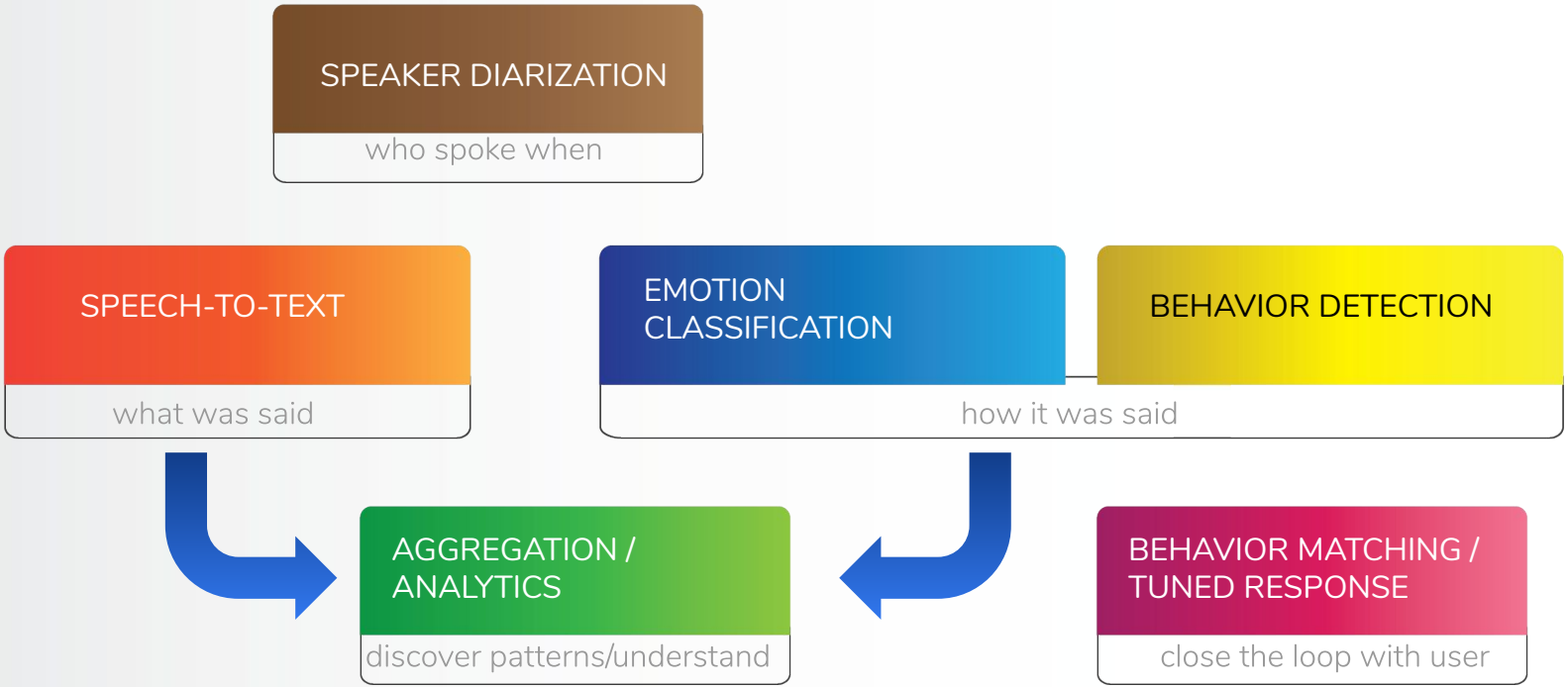
OLIVER API CAPABILITIES

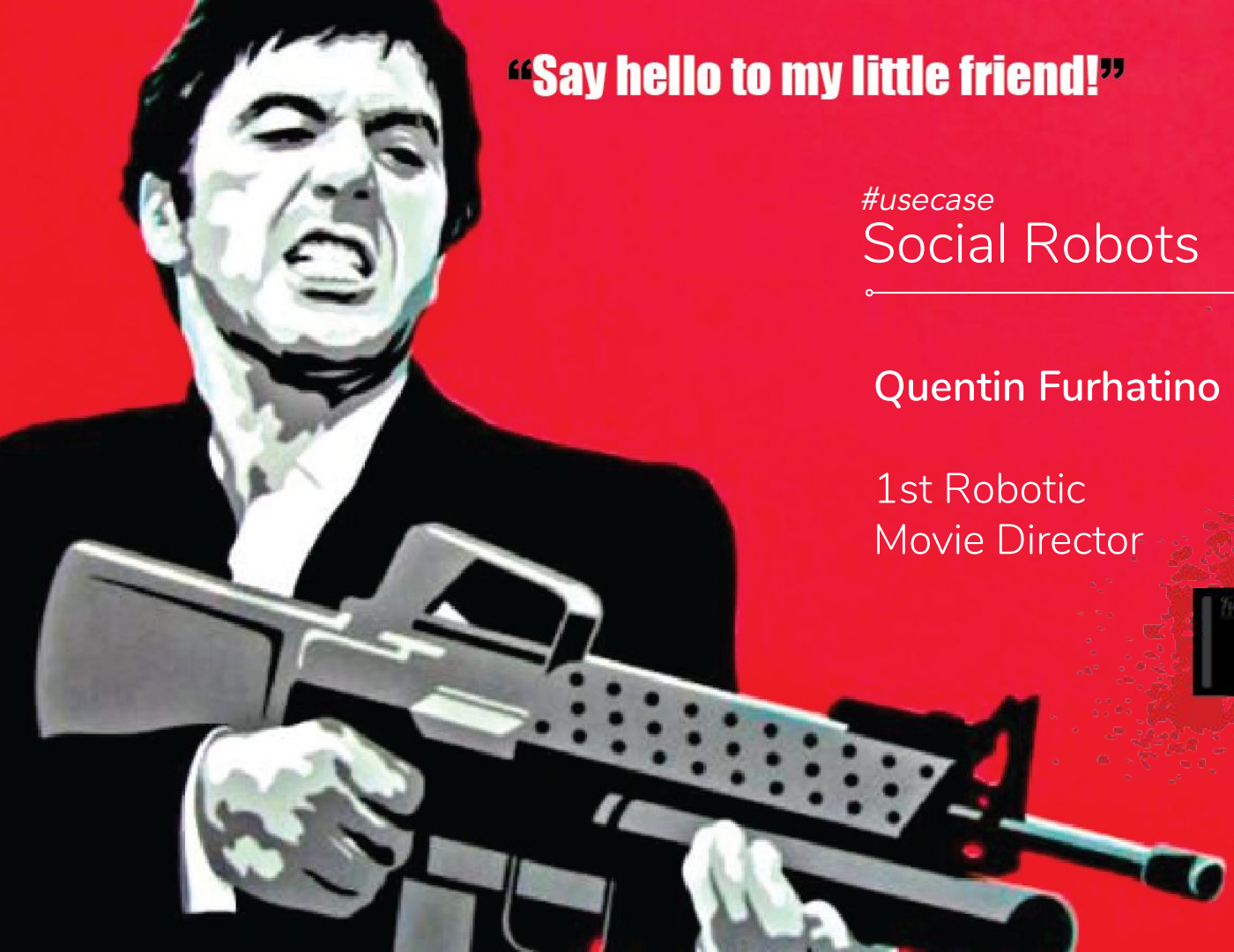
BEHAVIORAL SIGNALS

INTERACTION	+ Speaking rate + Overlap speech + Speaking ratio	+ Vocal variety + Active listening time
BEHAVIORAL	+ Positivity + Emotion (anger, happiness, sadness, neutral)	+ Engagement + Politeness + Agitation
KPIs	+ Interaction outcome prediction: e.g., successful?	+ Quality of interaction: engaging?

OLIVER API CAPABILITIES

BEHAVIORAL SIGNALS





Furhat ROBOTICS

#usecase

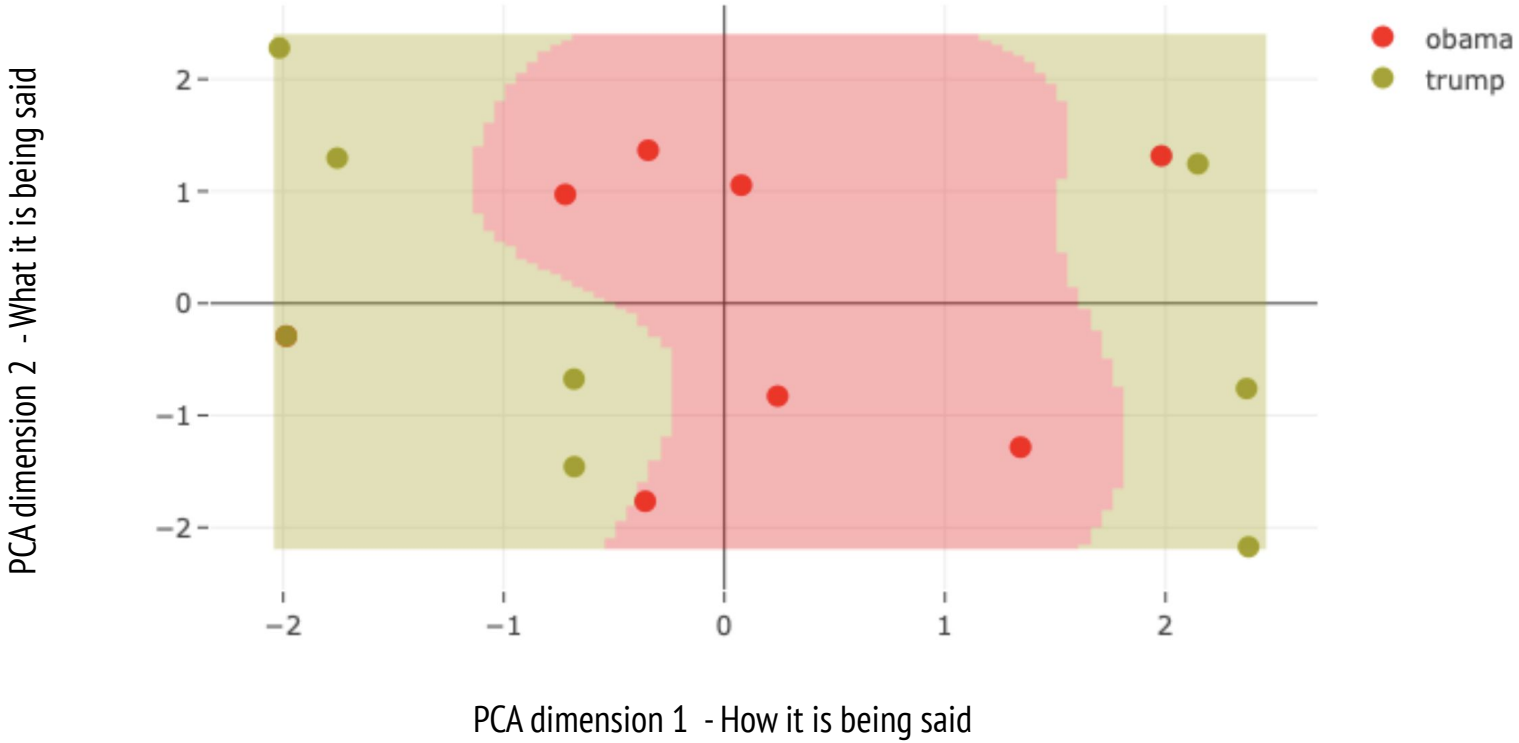
Social Robots

Quentin Furhatino

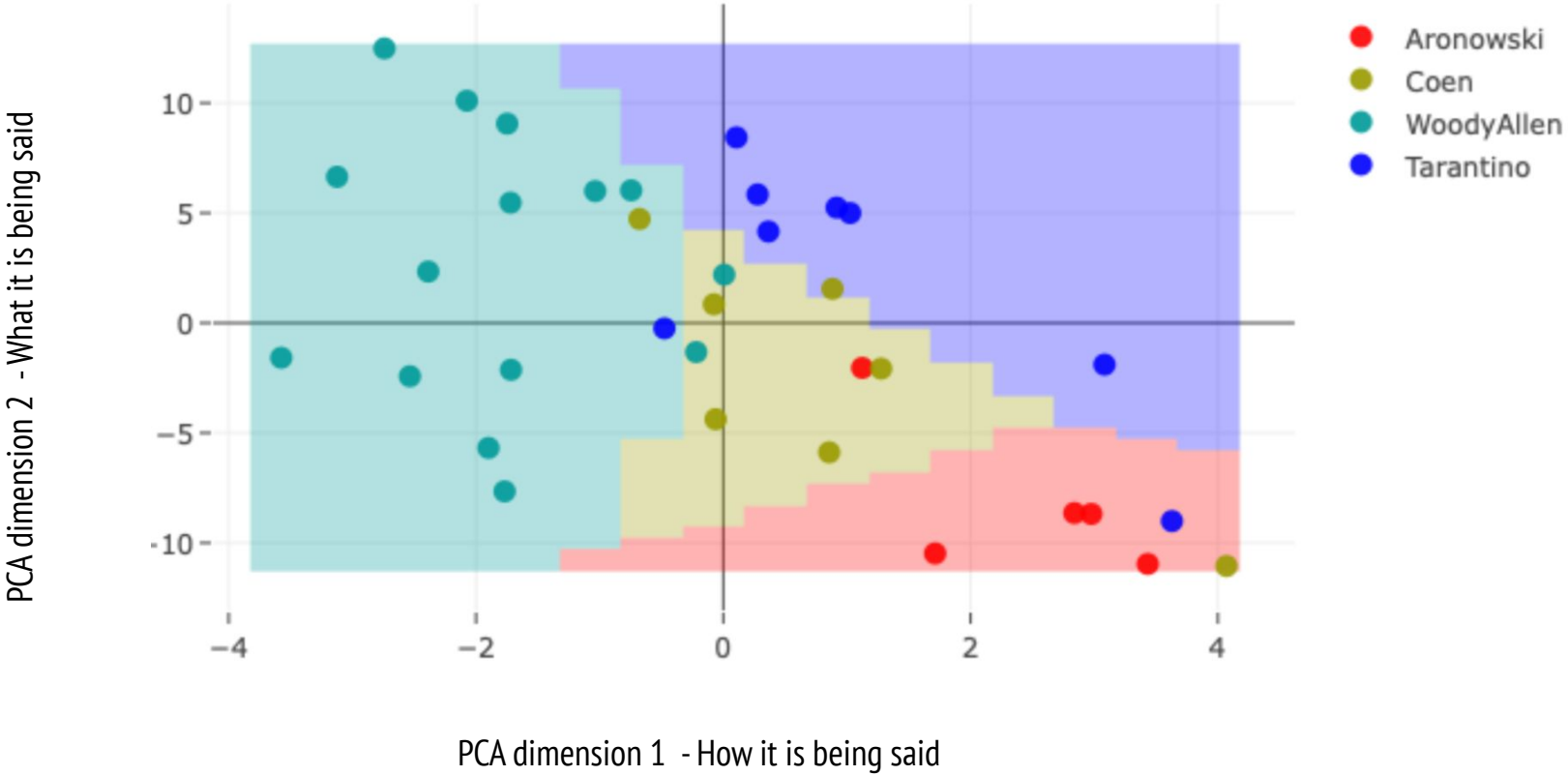
1st Robotic
Movie Director

BEHAVIORAL SIGNALS

Political Discourse Analysis



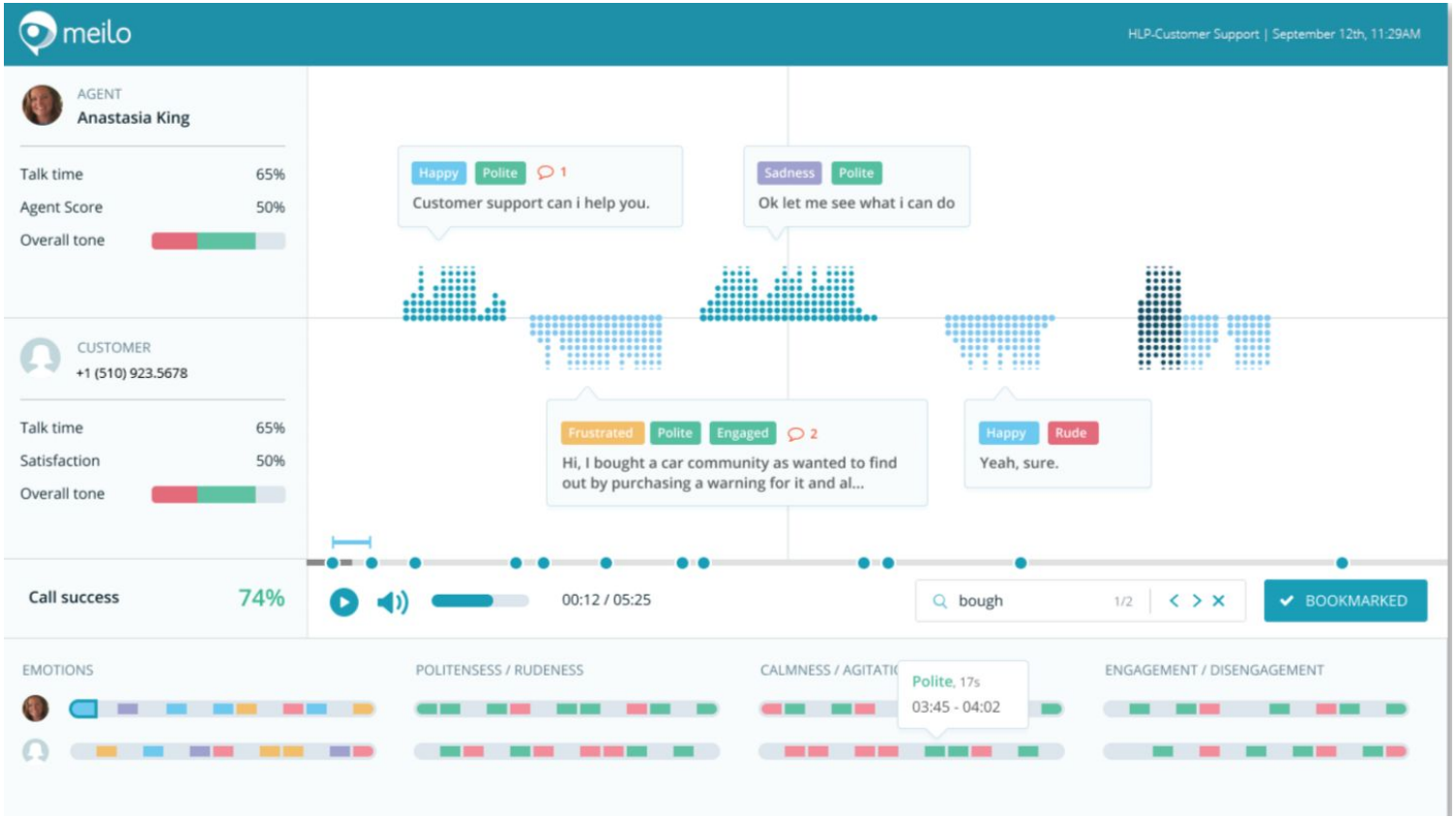
Media Analytics: Movies



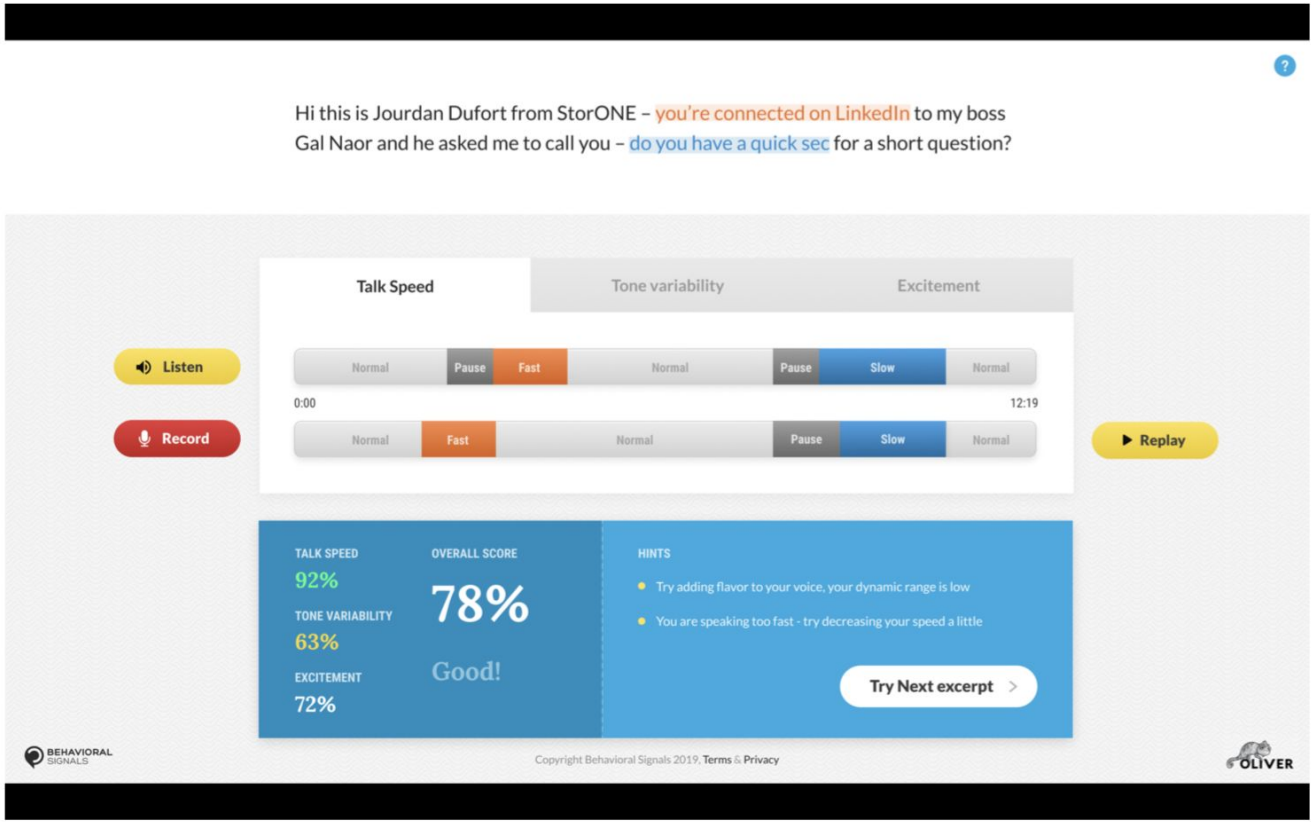
Emotionally Aware Subtitling

emosubs

Speech Analytics



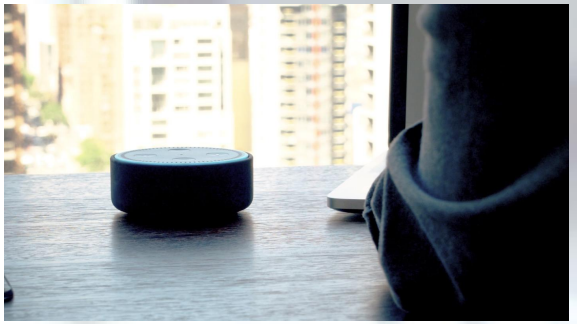
Speaker (Style) Training

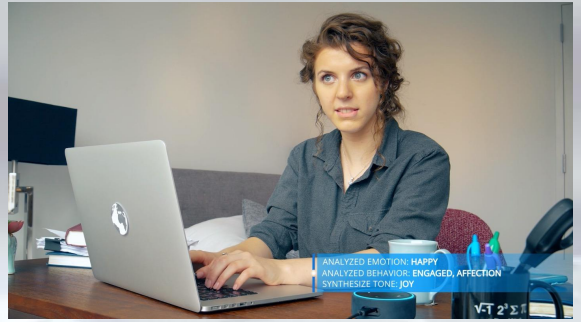


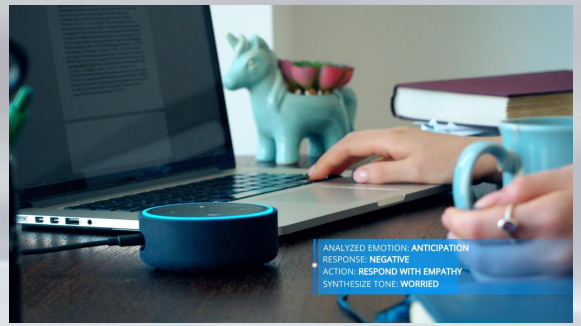
Emotionally Aware Gaming

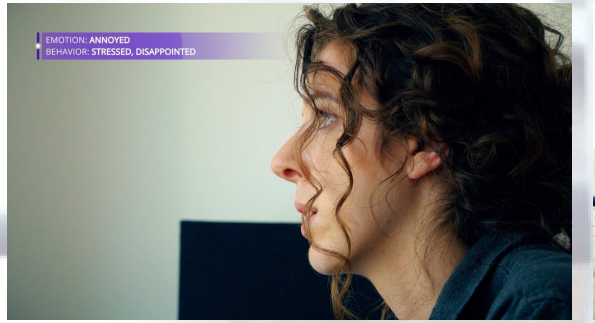
#usecase

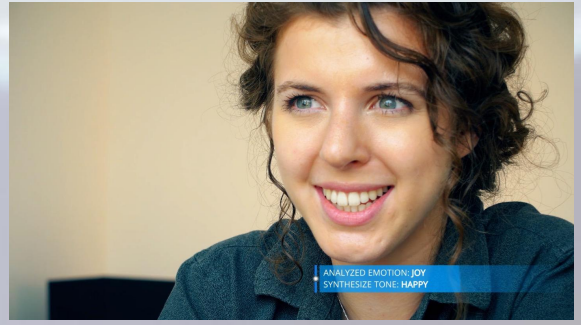
Virtual Assistant in Education











A virtual assistant that can support a student like a personal friend

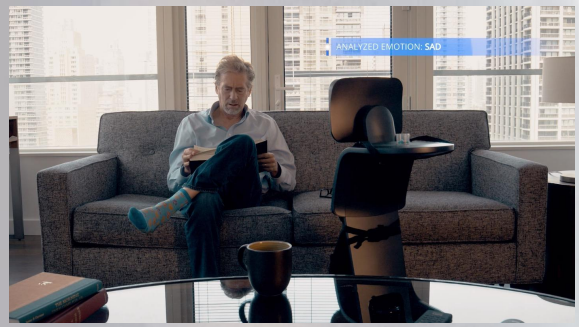
https://vimeo.com/309941043

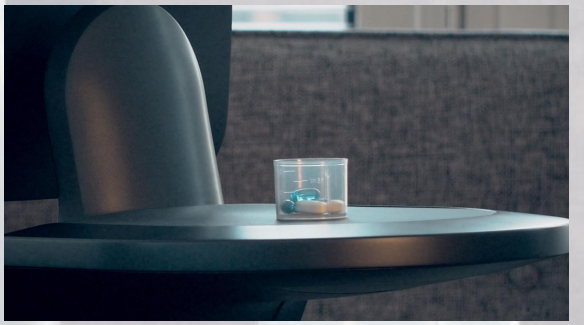
BEHAVIORAL SIGNALS

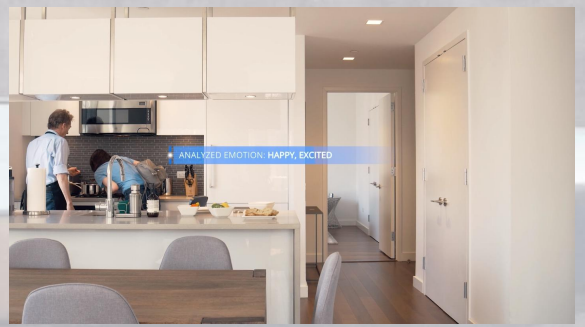
#usecase

Healthcare & Social Robots

temi
THE PERSONAL ROBOT







A social robot that can be a companion for those requiring care

https://vimeo.com/309927039

BEHAVIORAL SIGNALS

Why is it cool? Why is it hard?

"To summarize by paraphrasing Ekman — the inspiration behind the TV series *Lie to me* — **"emotions [and thoughts] determine the quality of our life"** while **behaviors and actions determine the outcomes in our life**."

"Human behavior offers a **window into the mind**. When we observe someone's actions, we are constantly **inferring his or her mental state** — their beliefs, intents, and knowledge — a concept known as theory of mind."

https://medium.com/behavioral-signals-ai/behavioral-signals-what-is-that-367ba0de49d2

BEHAVIORAL SIGNALS

Thanks for Listening

Q&A?