Manifold Learning & Nonlinear Recurrence Dynamics for Speech Emotion Recognition on Various Timescales

Efthymios Tzinis

School of Electrical and Computer Engineering (ECE),
National Technical University of Athens (NTUA), Greece

Introduction

- Develop computational systems capable of:
  - Hearing → Automatic Speech Recognition (ASR)
  - Feeling? → Speech Emotion Recognition (SER)
- ASR vs SER
  - ASR is about What you say?
  - SER is about How you say it?
- SER importance and applications:
  - Build adaptive Human Computer Interaction interfaces
  - Call centers, personal robot assistants, etc.
- How to build a SER system?
  - Extract representative sets of acoustic features
  - Select and train a classification model

Outline

1. Investigating how different timescales affect the performance of SER systems for:
   - Feature extraction
   - Model inference
2. Finding novel nonlinear acoustic features for SER
   - Exploiting recurrence dynamics of reconstructed phase spaces
   - Performance increment of SER systems based on conventional features
3. Developing a new algorithm for nonlinear dimensionality reduction
   - Derivative free optimization
   - Application to SER

Timescales of Emotional Inference

Idea Outline

- Assumption: SER performance depends on the timescale of emotion feature extraction
- Types of timescales of inferring emotional content
  - Frame ≈ 30 milliseconds
  - Phoneme ≈ 90 milliseconds
  - Speech segment $\approx 1 - 3$ seconds
  - Utterance
- How timescales affect SER performance for different:
  - Features $\rightarrow$ Extraction
  - Models $\rightarrow$ Inference

Feature Extraction Timescales: Local & Global (IS10 [23]) Features

- Selected Feature Sets (Left) & Statistical Functions (Right)
  - Low Level Descriptors (LLDs)

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

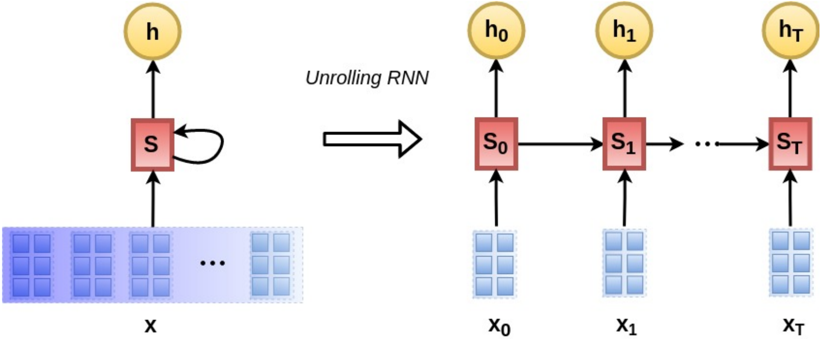
Statistical Functions	Set
position max/min arithmetic mean, standard deviation skewness, kurtosis linear regression coefficient 1/2 Quadratic & Absolute linear regression error quartile 1/2/3 quartile range 2-1/3-2/3-1 percentile 99 up-level time 75/90	A
percentile 1, percentile range 1-99	B
OnSets Number,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

[23] Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S., "The INTERSPEECH 2010 Paralinguistic Challenge," *INTERSPEECH*, pp. 2794–2797, 2010

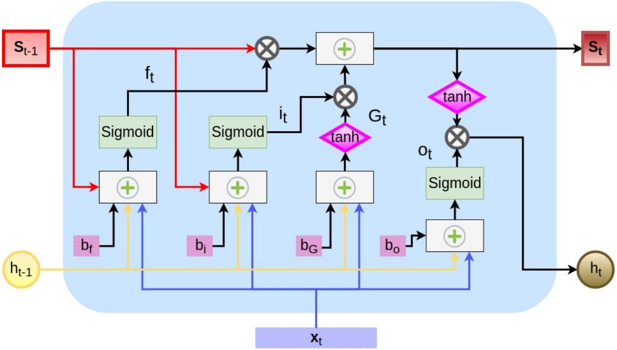
From Sub-Utterance Features to Utterance Inference

Recurrent Neural Network (RNN)

- Modeling sequences of vectors $\{\mathbf{x}_j\}_{j=0}^{T}$ (timesteps)
- Each timestep corresponds to a frame or a segment
- Multiple timesteps $\rightarrow$ backward gradient flow problem?



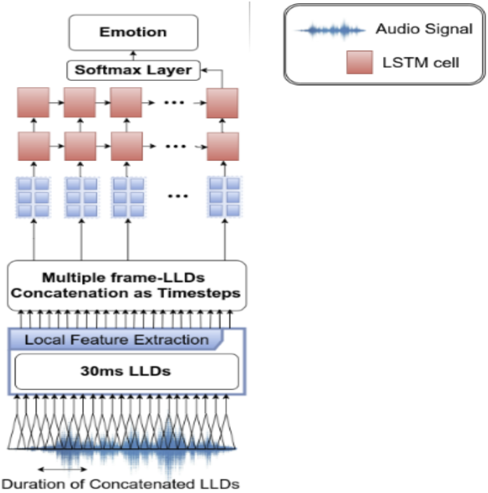
Long Short Term Memory (LSTM) unit



- Resistant to: Vanishing and exploding gradient
- Understanding longer time-dependencies

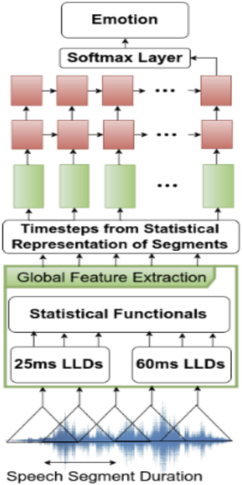
Submodule	Equation Update
Forget gate	$\mathbf{f}_t = \sigma(\mathbf{W}_f \cdot (\mathbf{h}_{t-1} \| \| \mathbf{x}_t) + \mathbf{b}_f)$
Input gate	$\mathbf{i}_t = \sigma(\mathbf{W}_i \cdot (\mathbf{h}_{t-1} \| \| \mathbf{x}_t) + \mathbf{b}_i)$ $\mathbf{G}_t = tanh(\mathbf{W}_G \cdot (\mathbf{h}_{t-1} \| \| \mathbf{x}_t) + \mathbf{b}_G)$
State	$\mathbf{S}_t = \mathbf{i}_t \odot \mathbf{G}_t + \mathbf{f}_t \odot \mathbf{S}_{t-1}$
Activation	$\mathbf{o}_t = tanh(\mathbf{W}_o \cdot (\mathbf{h}_{t-1} \| \| \mathbf{x}_t) + \mathbf{b}_o)$ $\mathbf{h}_t = \mathbf{o}_t \odot tanh(\mathbf{S}_t)$

Direct Approach using Frame-level Features



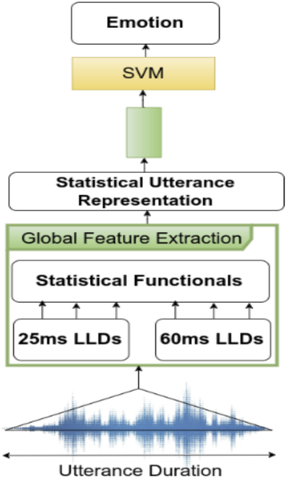
- Local Features: Frame-LLDs
- LSTM Trained with Concatenated Local Features

Segment-Based Approach using Global Features



- Global Features: Compute statistical functionals over extracted LLDs and create static-length representation
- LSTM Trained with Global Features over Segments (1582 features per segment)

Utterance-Based Approach using Global Features

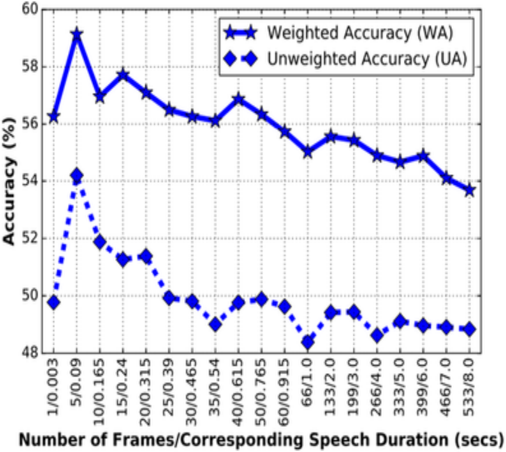


- Global Features: Compute statistical functionals over extracted LLDs and create static-length representation (1582 features per utterance)
- Support Vector Machine (SVM) Trained with Global Features over the whole Utterance

Investigating Timescales: Experimental Setup

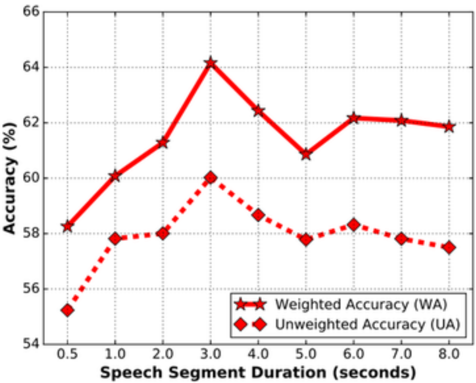
- Database: IEMOCAP
  - 5 Sessions: 2 speakers per session (1 Male, 1 Female)
  - 4490 emotional utterances
  - 4 Emotions: *Angry* (1103), *Sad* (1084), *Happy* (595), *Neutral* (1708)
- Evaluation Schema:
  - Leave One Session Out (LOSO): 5 folds (4 train, 1 test)
  - Test: 1 speaker for validation 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

Evaluation Results: LSTM with Local Features



- Best for phoneme timescale (5 frames - 90ms)
- By increasing the number of frames in every chunk there is a decline in performance

Evaluation Results: LSTM with Global Features



- Best for segments corresponding to 3 seconds
- Phoneme (0.5s) timescales do not contain sufficient emotional context
- Utterance (6-8s) time-scales reduce LSTM's expressiveness

Proposed Models Comparison with the Literature

Model	Type of Features	WA (%)	UA(%)
Best LSTM [35]	Spectrogram	61.71	58.05
BLSTM-SUA [66]	LLDs	59.33	49.96
BLSTM-WPA [65]	LLDs	63.5	58.8
BLSTM-ELM [61]	LLDs chunks of 250ms	62.85	63.89

Model	Type of Features	WA (%)	UA(%)
SVM	IS10 over the whole utterance	53.54	49.23
LSTM	LLDs chunks of 90ms	59.14	54.2
LSTM	IS10 over 3 seconds segments	64.16	60.02

- State-of-the-art results on IEMOCAP using a simple LSTM

[35] Fayek, H., M., Lech, M. and Cavedon, L., (in press), "Evaluating deep learning architectures for Speech Emotion Recognition," Neural Networks, vol. 92, pp. 60–68, 2017.
[66] Huang, C., W., Narayanan, S., "Attention Assisted Discovery of SubUtterance Structure in Speech Emotion Recognition," in Proceedings of INTERSPEECH, 2016, pp. 1387–1391.
[65] Mirsamadi, S., Barsoum, E. and Zhang, C., (in press), "Automatic Speech Emotion Recognition Using Recurrent Neural Networks With Local Attention," in Proceedings of ICASSP, 2017, pp. 2227–2231.
[61] Lee, J. and Tashev, I., "High-level feature representation using recurrent neural network for speech emotion

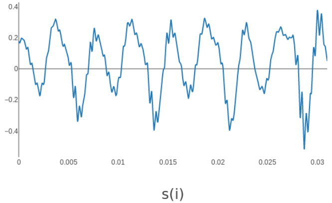
Integrating Nonlinear Recurrence Dynamics

Motivation

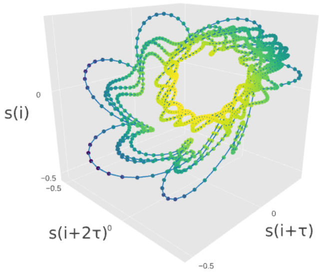
- Linearity Assumptions in Voice Modeling:
  - Short-term speech signals (≈ 30 ms) are stationary
    - Everything that uses a Fourier transformation
    - Mel Frequency Cepstral Coefficients (MFCCs), etc.
  - Linear Predictive Coding (LPC)
- Too good to be true
- The process of speech production is generally nonlinear
  - Modulations of the speech airflow and turbulence
  - Biphonation (two independent pitches)
  - Nonlinear laryngeal vibrations

Recurrence Properties of Speech

- Recurrence properties of speech dynamics
  - Reconstruct the phase space of each speech frame
  - Recurrence structures from the co-evolution of trajectories
  - Integrate the emerging recurrence patterns?



Time-domain representation of the speech frame



Reconstructed Phase Space

Reconstruction of the Phase Space (PS)

- Idea (*Intelligible Realm, Phaedrus by Plato* ≈ 370 *BC*):
  - Observed signal s is only a projection of the true signal $\mathrm{s}^{\star}$
  - Approximate PS using time-delayed versions of s
- Definition of the PS trajectory:

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

- Estimate τ using Average Mutual Information (AMI):

$$\mathcal{I}(\mathbf{s}, \tau) = \sum_{i=1}^{N-\tau} p_b(s(i), s(i+\tau)) \cdot log_2[\frac{p_b(s(i), s(i+\tau))}{p_b(s(i)) \cdot p_b(s(i+\tau))}]$$

- Estimate d_e using False Nearest Neighbors (FNN):

$$R_{FNN}^{\hat{d}_e}(\mathbf{x}(i), \mathbf{x}(j)) = \frac{\mathbf{D}_{\hat{d}_e+1}(\mathbf{x}(i), \mathbf{x}(j)) - \mathbf{D}_{\hat{d}_e}(\mathbf{x}(i), \mathbf{x}(j))}{\mathbf{D}_{\hat{d}_e}(\mathbf{x}(i), \mathbf{x}(j))}$$

Computation of Recurrent Plots (RPs)

- RP: Thresholded distance matrix from PS orbits $\{\mathbf{x}(i)\}_{i=1}^{N}$

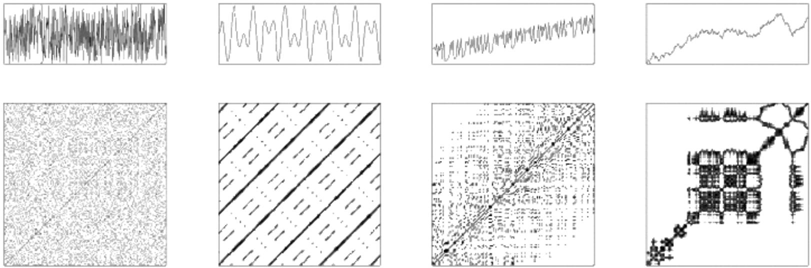
$$\mathbf{R}_{i,j}(\epsilon, q) = \Theta(\epsilon - ||\mathbf{x}(i) - \mathbf{x}(j)||_q)$$

- Setting threshold parameter ϵ:
  1. Ad-hoc selection
  2. Based on stabilizing recurrence density
  3. Based on a fixed ratio of the standard deviation of points
- Setting norm parameter q:
  - Manhattan: $q = 1$, Euclidean: $q = 2$, Supremum: $q = \infty$
- Recurrence structures are based on points and lines
- Example of an L-length diagonal line (of ones):

$$(1 - \mathbf{R}_{i-1,j-1})(1 - \mathbf{R}_{i+L+1,j+L+1}) \prod_{k=1}^{k=L} \mathbf{R}_{i+k,j+k} = 1$$

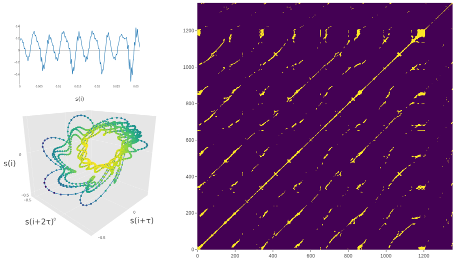
What's Special About These Visualizations?

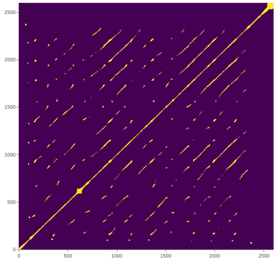
- RPs can visualize the identity of the underlying dynamics!
  - They have not yet been utilized for SER



Recurrence plots for different types of systems. From left to right: random noise, periodic oscillations with two frequencies, deterministic chaotic system and autoregressive process.
https://en.wikipedia.org/wiki/Recurrence_plot

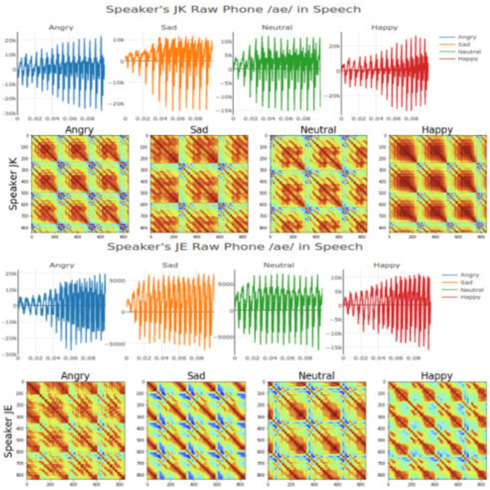
Analysis of Speech Dynamics using RPs





(Left): RP of a 30ms frame contained in the excitation of vowel /e/ inside an angry utterance
(Right): RP of Lorenz96 system displaying chaotic behavior

Intuition behind RPs for Emotional Speech



- Pitch-periodic motifs ✓
- Single isolated points → strong fluctuation ✗
- Small diagonal lines → chaos dynamics ✓
- Some bowed lines → changing of dynamics ✓
- Vertical-horizontal lines → laminar states ✓
- White bands → nonstationary data (rare states) ✓

Recurrence Quantification Analysis (RQA) Feature Set

RQA Measure	Formulation
Recurrence Rate	$\frac{1}{N^2}\sum_{i,j=1}^{N}\mathbf{R}_{i,j}$
Determinism	$\frac{\sum_{l=d_m}^{N} lP_d(l)}{\sum_{l=d_m}^{N} lP_d(l)}$
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	$\frac{\sum_{l=v_m}^{N} lP_v(l)}{\sum_{l=v_m}^{N} lP_v(l)}$
Max Vertical Length	$max(\{l_i\}_{i=1}^{N_v})$
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	$max(\{l_i\}_{i=1}^{N_w})$
Average White Vertical Length	$\frac{\sum_{l=w_m}^{N} lP_w(l)}{\sum_{l=w_m}^{N} P_w(l)}$
White Vertical Entropy	$\sum_{l=w_m}^{N}\frac{P_w(l)}{N_w}ln(\frac{N_w}{P_w(l)})$

Statistical Functions
min
max
arithmetic mean
median
variance
skewness
kurtosis
range
$1_{st}, 5_{th}, 25_{th}, 50_{th}, 75_{th}, 95_{th}, 99_{th}$ percentiles
$25-50$, $50-75$ and $25-75$ quartile ranges

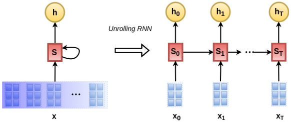
- 432 Features

Experimental Setup

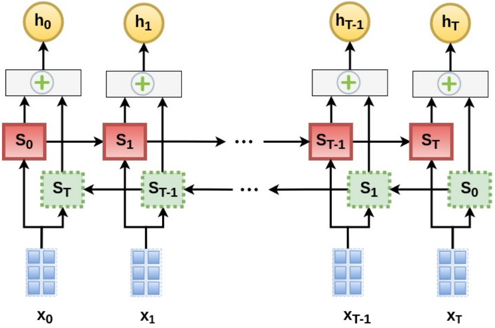
- Datasets
  1. Surrey Audio-Visual Expressed Emotion (SAVEE)
     - 480 utterances, 7 emotions
  2. Berlin Database of Emotional Speech (Emo-DB)
     - 535 utterances in German, 7 emotions
  3. IEMOCAP
     - 5 Sessions (2 speakers each), 4 Emotions, 5531 utterances (*Angry*, *Sad*, *Happy + Excited*, *Neutral*)
- Approaches on different timescales
  - Utterance-based: SVM and Logistic Regression (LR)
  - Segment-based: Attention-Bidirectional LSTM (A-BLSTM) (1 second segments, 0.5 overlap)
- Evaluated feature sets:
  - (Proposed) RQA: 432 features
  - IS10: 1582 features
  - (RQA + IS10): 2014 features

Bidirectional LSTM (BLSTM)

- Using opposite time-direction flows for $\{\mathbf{x}_j\}_{j=0}^{T}$ (timesteps)
  - Concatenate activations $\mathbf{h}_t = \overrightarrow{\mathbf{h}}_t || \overleftarrow{\mathbf{h}}_{T-t}$



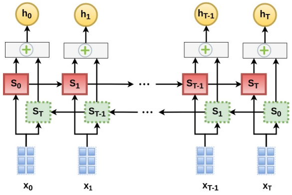
Simple RNN



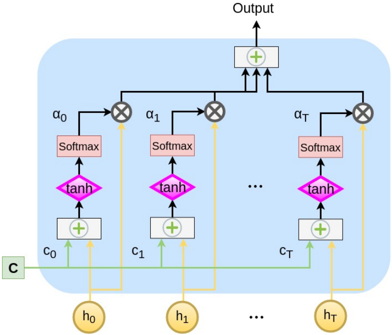
Bidirectional RNN

Attention-based BLSTM (A-BLSTM)

- Focusing only on the most important timesteps of $\{\mathbf{x}_j\}_{j=0}^{T}$
  - Trainable and normalized attention vector $\mathbf{a}$
- Output: $\sum_{t=0}^{T} a_t \odot \mathbf{h}_t$
- $\sum_{t=0}^{T} a_t = 1$



BLSTM



Attention Mechanism

Speaker Dependent (SD) Experiments

Features	Model	SAVEE WA	SAVEE UA	Emo-DB WA	Emo-DB 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
[37] Spectrogram	SAE	75.4	-	88.3	-
[49] LLDs Stats	ESR	76.3	73.4	88.7	87.9

- Utterance-based
- (PS-N) Per-Speaker z-Normalization
- 5-fold cross-validation

- (RQA+IS10) set yields improvement compared to IS10:
  - 3.1 % in WA and 3.4 % in UA for SAVEE
  - 4.9 % in WA and 5.7 % in UA for Emo-DB
- Improvement compared to models in the literature of up to:
  - 4.8 % in WA and 5.0 % in UA for SAVEE
  - 5.0 % in WA and 4.5 % in UA for Emo-DB

[37] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, “Learning salient features for speech emotion recognition using convolutional neural networks,ÏEEE Transactions on Multimedia, vol. 16, no. 8, pp. 2203–2213, 2014.
[49] Y. Sun and G. Wen, “Ensemble softmax regression model for speech emotion recognition,”Multimedia Tools and Applications, vol. 76, no. 6, pp. 8305–8328, 2017.

Speaker Independent (SI) Experiments

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
[49] LLDs Stats	ESR	51.5	49.3	**82.4**	**78.7**
[143] WSFHM+IS10	SVM	50.0	-	81.7	-

- Utterance-based
- (PF-N) Per-Fold z-Normalization
- One-speaker-out cross-validation

- (RQA+IS10) set yields improvement compared to IS10:
  - 5.5 % in WA and 8.2 % in UA for SAVEE
  - 2.4 % in WA and 3.2 % in UA for Emo-DB
- Improvement compared to models in the literature of up to:
  - 4.0 % in WA and 4.5 % in UA for SAVEE
  - 0.4 % in WA for Emo-DB

[49] Y. Sun and G. Wen, “Ensemble softmax regression model for speech emotion recognition,”Multimedia Tools and Applications, vol. 76, no. 6, pp. 8305–8328, 2017.
[143] Y. Sun, G. Wen, and J. Wang, “Weighted spectral features based on local hu moments for speech emotion recognition,”Biomedical signal processing and control, vol. 18, pp. 80–90, 2015.

LOSO Experiments on IEMOCAP

Features	Model	PS-N		PF-N		G-N	
		WA	UA	WA	UA	WA	UA
IS10	SVM	58.3	60.9	58.9	60.1	59.2	60.5
	LR	57.5	61.2	54.6	57.9	53.5	57.5
	A-BLSTM	62.0	65.1	62.6	65.0	62.8	65.0
RQA	SVM	52.9	54.6	53.1	53.8	53.1	53.7
	LR	52.2	54.8	52.6	54.0	52.8	54.3
	A-BLSTM	55.6	59.3	56.6	58.3	56.7	58.7
RQA + IS10	SVM	59.3	61.8	59.2	60.4	59.5	60.7
	LR	58.3	62.0	55.6	58.7	54.5	58.7
	A-BLSTM	**62.7**	**65.8**	**63.0**	**65.2**	62.9	**65.5**
[27] MFB	CNN	-	61.8	-	-	-	-
[28] IS10	DBN	-	-	-	-	60.9	62.4
[35] SP	CNN	-	-	-	-	**64.8**	60.9
[36] GFS	BLSTM	-	-	50.5	51.9	-	-

- Consistent improvement using the fused set compared to IS10
- State-of-the-art on IEMOCAP

[27] Z. Aldeneh and E. M. Provost, “Using regional saliency for speech emotion recognition,ïn Proceedings of ICASSP, 2017, pp. 2741–2745.
[28] R. Xia and Y. Liu, “A multi-task learning framework for emotion recognition using 2d continuous space,ÏEEE Transactions on Affective Computing, vol. 8, no. 1, pp. 3–14, 2017.
[35] H. M. Fayek, M. Lech, and L. Cavedon, “Evaluating deep learning architectures for speech emotion recognition,”Neural Networks, vol. 92, pp. 60–68, 2017.
[36] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, “Representation learning for speech emotion recognition.ïn Proceedings of INTERSPEECH, 2016, pp. 3603–3607.

Pattern Search MDS

Dimensionality Reduction

- Huge dimensionality of N feature vectors
  - High-dimensional representations $\mathbf{Y} \in \mathbb{R}^{N \times D}$
  - Are all these features mandatory for an apt representation?
- Could we find a lower dimensional space or manifold embedded in this space $\mathbf{X} \in \mathbb{R}^{N \times L}$, where $L \ll D$?
  - Preserve the geometry of the given data $\mathbf{Y} \in \mathbb{R}^{N \times D}$
  - Producing competitive classification accuracies for SER
- Why?
  - $\searrow$ Training time
  - $\nearrow$ Accuracy ? (Curse of Dimensionality)
  - Visualization

Multidimensional Scaling (MDS)

- Multidimensional Scaling (MDS)
  - Searching for a solution preserving the pairwise distances of the high dimensional space, e.g., $d_{ij}(\mathbf{X}) \approx d_{ij}(\mathbf{Y})$
  - Minimizing Stress:
    $$\sigma^2_{raw}(\mathbf{D}_\mathbf{X}, \mathbf{D}_\mathbf{Y}) = \sum_{i,j} w_{ij}[d_{ij}(\mathbf{X}) - d_{ij}(\mathbf{Y})]^2$$
  - Could be extended to geodesic distances
  - Until now: Iterative algorithms based on gradient descent or minimizing a majorization convex function, e.g., SMACOF
- Gradient-free MDS?
  - Better solutions
  - Faster convergence
  - Proof of convergence
  - Application on SER

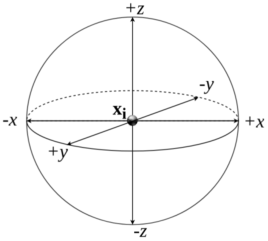
Pattern Search MDS

procedure MDS($\mathbf{D}_\mathbf{Y}$, L, $r^{(0)}$)
  $k \leftarrow 0$
  $\mathbf{X}^{(k)} \leftarrow$ UNIFORM($N \times L$)
  $\mathbf{D}^{(k)} \leftarrow$ DISTANCE_MATRIX($\mathbf{X}^{(k)}$)
  $e^{(k)} \leftarrow f(\mathbf{D}_\mathbf{Y}, \mathbf{D}^{(k)})$
  $e^{(k-1)} \leftarrow +\infty$
  $r^{(k)} \leftarrow r^{(0)}$
  while $r^{(k)} > \delta$ **do**
    if $e^{(k-1)} - e^{(k)} \leq \epsilon \cdot e^{(k)}$ **then**
      $r^{(k)} \leftarrow \frac{r^{(k)}}{2}$
    $\mathbf{S} \leftarrow$ SEARCH_DIRECTIONS($r^{(k)}$, L)
    for all $x \in \mathbf{X}^{(k)}$ **do**
      $\mathbf{X}^*, e^* \leftarrow$ OPTIMAL_MOVE($\mathbf{X}^{(k)}$,x,$\mathbf{S}$,$e^{(k)}$)
      $e^{(k-1)} \leftarrow e^{(k)}$
      $e^{(k)} \leftarrow e^*$
      $\mathbf{X}^{(k)} \leftarrow \mathbf{X}^*$
    $k = k + 1$

- Target distance matrix $\mathbf{D}_\mathbf{Y}$
- Target embedding dimension L
- Iteration index k
- $\mathbf{D}^{(k)} = d_{ij}(\mathbf{X}^{(k)})$
- $e^{(k)} = \sigma_{raw}(\mathbf{D}_\mathbf{Y}, \mathbf{D}^{(k)})$
- Search radius $r^{(k)}$
- Search moves independently for each point $\mathbf{x}_i \in \mathbf{X}^{(k)}$

Search Directions and Optimal Moves

function SEARCH_DIRECTIONS(r, L)
$\quad \mathbf{S}^{+} \leftarrow r \cdot \mathbf{I}_L$
$\quad \mathbf{S}^{-} \leftarrow -r \cdot \mathbf{I}_L$
$\quad$ **return** $\mathbf{S}^{+} || \mathbf{S}^{-}$



function OPTIMAL_MOVE($\mathbf{X}^{(k)}$, x, S, e)
$\quad e^* \leftarrow e$
$\quad$ **for all** $s \in \mathbf{S}$ **do**
$\quad\quad \tilde{x} \leftarrow x + s$
$\quad\quad \tilde{\mathbf{X}} \leftarrow$ UPDATE_POINT($\mathbf{X}^{(k)}$, x, $\tilde{x}$)
$\quad\quad \mathbf{D} \leftarrow$ DISTANCE_MATRIX($\tilde{\mathbf{X}}$)
$\quad\quad \tilde{e} \leftarrow \sigma_{raw}(\mathbf{D_Y}, \mathbf{D})$
$\quad\quad$ **if** $\tilde{e} < e^*$ **then**
$\quad\quad\quad e^*, \mathbf{X}^* \leftarrow \tilde{e}, \tilde{\mathbf{X}}$
$\quad$ **return** X^*, e^*

- Search over Cartesian coordinates
- Move along optimal s if it reduces the loss (min $\sigma_{raw}(\mathbf{D_Y}, \mathbf{D}^{(k)})$)
- Else do not move that point
- Complexity $\mathcal{O}(N^2 L)$ per epoch

General Pattern Search (GPS) Methods

```
procedure GPS_SOLVER(x^(0), Δ^(0), C^(0))
    k = −1
    do
        k = k + 1
        s^(k) = EXPLORE(P^(k), x^(k), Δ^(k))
        ρ^(k) = f(x^(k) + s^(k)) − f(x^(k))
        if ρ^(k) < 0 then
            x^(k+1) = x^(k) + s^(k)
        else
            x^(k+1) = x^(k)
        Δ^(k+1) = UPDATE(Δ^(k), ρ^(k))
        C^(k+1) = UPDATE(C^(k), ρ^(k))
    while convergence criterion == False
```

- Goal: Minimize $f : \mathbb{R}^n \to \mathbb{R}$
- Solution: $\mathbf{x}^* = \underset{\mathbf{x} \in \mathbb{R}^n}{\operatorname{argmin}} f(\mathbf{x})$
- Nonsingular basis $\mathbf{B} \in \mathbb{R}^{n \times n}$
- Generating matrix: $\mathbf{C}^{(k)} = [\mathbf{\Psi}^{(k)} \;\; \mathbf{L}^{(k)}]$
- $\mathbf{\Psi}^{(k)} = [\mathbf{M}^{(k)} \;\; -\mathbf{M}^{(k)}]$, $\mathbf{M}^{(k)} \in \mathbb{Z}^{n \times n}$
- $\mathbf{0} \in \mathbf{L}^{(k)}$ (non-movement)
- Pattern matrix: $\mathbf{P}^{(k)} = \mathbf{B}\mathbf{C}^{(k)}$
- Step-length parameter: $\Delta^{(k)}$
- Trial Move: $\mathbf{s}_i^{(k)} = \Delta^{(k)} \mathbf{p}_i^{(k)}$
- Unsuccessful iterations: $\Delta^{(k+1)} < \Delta^{(k)}$
- Successful iterations: $\Delta^{(k+1)} \geq \Delta^{(k)}$

Pattern Search MDS Expressed as GPS Instance

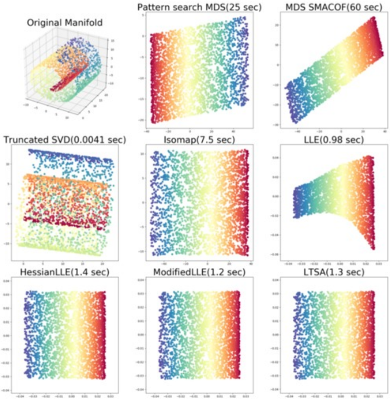
Name	GPS Formulation	Pattern Search MDS Formulation
Variable and Search Space	$\mathbf{x} \in \mathbb{R}^n$	$\mathbf{z} = vec(\mathbf{X}^T) \in \mathbb{R}^{N \cdot L}$ $\mathbf{z} = [x_{11}, ..., x_{1L}, ..., x_{N1}, ..., x_{NL}]^T$
Objective Function	$f : \mathbb{R}^n \to \mathbb{R}$	$g(\mathbf{z}) = \sum_{i,j} (d_{ij}(\mathbf{z}) - d_{ij}(\mathbf{D_Y}))^2$
Solution	$\mathbf{x}^* = \underset{\mathbf{x} \in \mathbb{R}^n}{\text{argmin}}\ f(\mathbf{x})$	$\mathbf{z}^* = \underset{\mathbf{z} \in \mathbb{R}^{N \cdot L}}{\text{mín}}\ g(\mathbf{z})$
Nonsingular basis	$\mathbf{B}$	$\hat{\mathbf{B}} = \mathbf{I}_{N \cdot L} = [\mathbf{e}_1, ..., \mathbf{e}_{N \cdot L}]$
Generator Matrix	$\mathbf{C}^{(k)}$	$\hat{\mathbf{C}} = [\mathbf{I}_{N \cdot L} \quad -\mathbf{I}_{N \cdot L} \quad \mathbf{0}]$
Pattern Matrix	$\mathbf{P}^{(k)}$	$\hat{\mathbf{P}} \equiv \hat{\mathbf{B}}\hat{\mathbf{C}} \equiv \hat{\mathbf{C}}$
Step-length parameter	$\Delta^{(k)}$	Search Radius $r^{(k)}$
Trial Move	$\mathbf{s}_i^{(k)} = \Delta^{(k)} \mathbf{p}_i^{(k)}$	$\mathbf{s}_i^{(k)} = r^{(k)} \hat{\mathbf{p}}_i^{(k)}$
Unsuccessful iterations	$\Delta^{(k+1)} < \Delta^{(k)}$	$r^{(k+1)} = \frac{1}{2} r^{(k)}$
Successful iterations	$\Delta^{(k+1)} \geq \Delta^{(k)}$	$r^{(k+1)} = r^{(k)}$

Proof of Convergence

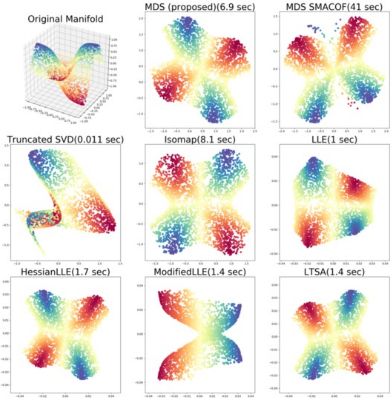
- If the following hold then: $\lim_{k \to +\infty} \inf \; ||\nabla f(\mathbf{x}^{(k)})|| = 0$
  1. Let $L(\mathbf{x}^{(0)}) = \{\mathbf{x} : f(\mathbf{x}) \leq f(\mathbf{x}^{(0)})\}$ be closed and bounded, f is continuously differentiable on the union of the open balls $\bigcup_{\mathbf{a} \in L(\mathbf{x}^*)} B(\mathbf{a}, \eta)$
  2. $\mathbf{s}_i^{(k)} = \Delta^{(k)} \mathbf{p}_i^{(k)} = \Delta^{(k)} \mathbf{B}[\mathbf{M}^{(k)} \;\; -\mathbf{M}^{(k)} \;\; \mathbf{L}^{(k)}]$
  3. If among the exploratory moves $\mathbf{a}^{(k)}$ at iteration k selected from the columns of the matrix $\Delta^{(k)} \mathbf{B}[\mathbf{M}^{(k)} - \mathbf{M}^{(k)}]$ exist at least one move that leads to success, i.e., $f(\mathbf{x}^{(k)} + \mathbf{a}) < f(\mathbf{x}^{(k)})$, then the EXPLORE_MOVES() subroutine will return a move $\mathbf{s}^{(k)}$ such that $f(\mathbf{x}^{(k)} + \mathbf{s}^{(k)}) < f(\mathbf{x}^{(k)})$.
- Indeed Pattern Search MDS converges to a fixed point
  1. Stress function is continuously differentiable everywhere on the union of open balls except of the edge case where $\mathbf{x}_i = \mathbf{x}_j$ [177]
  2. $\mathbf{s}_i^{(k)} = r^{(k)} \hat{\mathbf{p}}_i^{(k)}$
  3. Each epoch searches over all columns of $\hat{\mathbf{\Psi}} = [\mathbf{I}_{N \cdot L} \;\; -\mathbf{I}_{N \cdot L}]$

[177] V. Torczon, “On the convergence of the multidirectional search algorithm,”*SIAM journal on Optimization*, vol. 1, no. 1, pp. 123–145, 1991.

Manifold Geometry Reconstruction

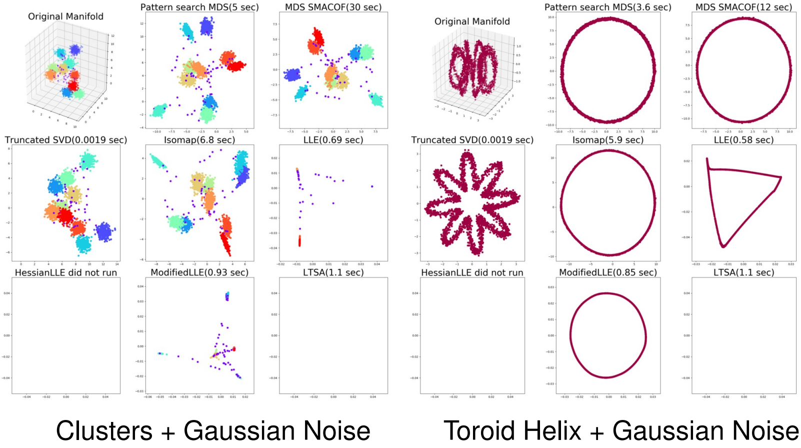


Swissroll

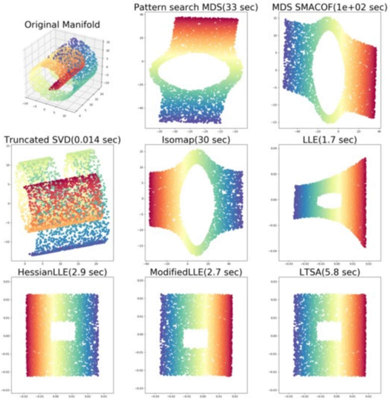


Twin Peaks

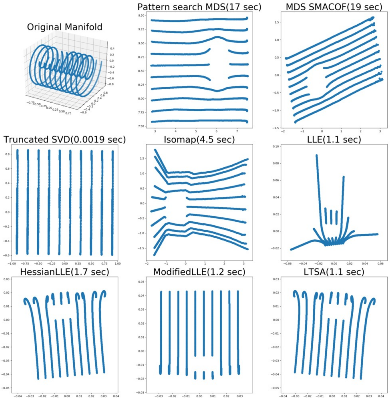
Robustness to Noisy Data



Robustness to Missing Data

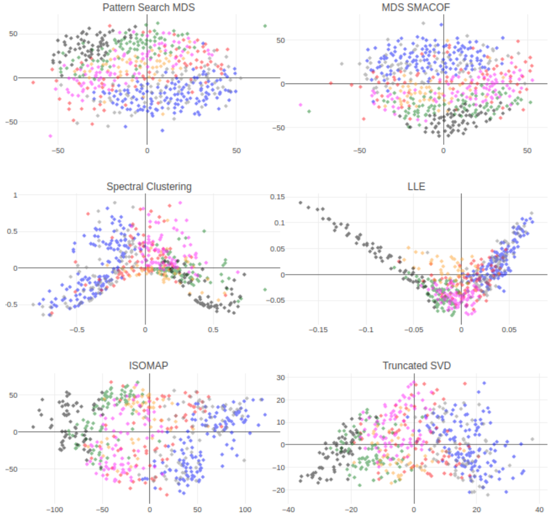


Swisshole



Spiral Hole

SER on Emo-DB with Reduced IS10 and KNN



- 1582 Features
- SMACOF
- 17-NN
- $L = 25$
- WA: 71.4 %
- UA: 65.5 %

SER on Emo-DB with Reduced RQA and KNN

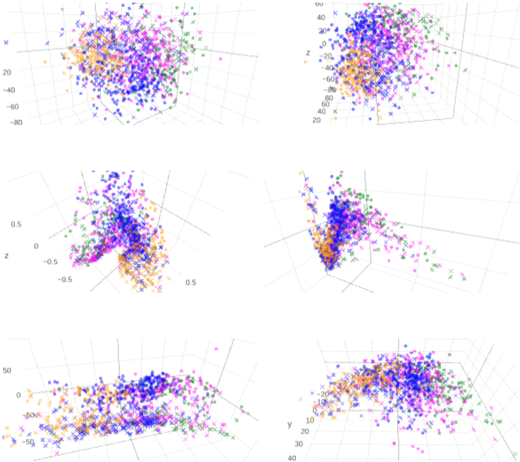
- 432 Features
- Pattern Search MDS
- 13-NN
- $L = 10$
- WA: 60.0 %
- UA: 52.7 %

SER on Emo-DB with Reduced (RQA+IS10) and KNN

- 2014 Features
- Pattern Search MDS
- 17-NN
- $L = 25$
- WA: 74.4 %
- UA: 68.8 %

3D Embeddings from 2 IEMOCAP Speakers

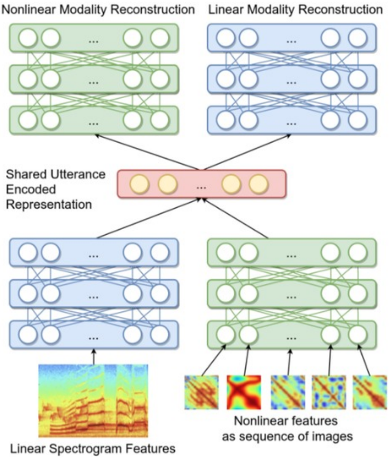
From left to right: Pattern Search MDS, MDS SMACOF, Spectral Clustering, LLE, ISOMAP, Truncated SVD



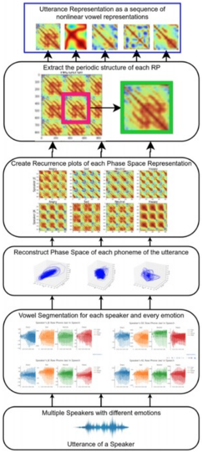
Comparison on Utterance Level SER

Features	Dimensionality Reduction Method	L	Classifier	EmoDB		IEMOCAP	
				WA	UA	WA	UA
IS10	-	1582	SVM	79.7	74.3	59.2	60.5
	-	1582	LR	76.1	71.9	53.5	57.5
	-	1582	KNN	69.9	64.1	53.1	55.7
	Pattern Search MDS	10	KNN	65.7	59.9	53.8	55.2
	Pattern Search MDS	25	KNN	70.4	63.5	**54.5**	56.8
RQA	-	432	SVM	70.9	64.2	53.1	53.7
	-	432	LR	71.1	67.1	52.8	54.3
	-	432	KNN	56.9	48.4	46.9	48.8
	Pattern Search MDS	10	KNN	60.0	52.7	46.4	47.2
	Pattern Search MDS	25	KNN	58.8	50.9	47.6	49.3
RQA+IS10	-	2014	SVM	82.1	76.9	59.5	60.7
	-	2014	LR	80.1	77.5	54.5	58.7
	-	2014	KNN	72.4	65.9	52.6	55.1
	Pattern Search MDS	10	KNN	69.9	63.1	52.9	54.4
	Pattern Search MDS	25	KNN	74.4	68.8	**54.9**	57.2

Future Work



Bimodal Autoencoder



Extract Periodic RP Motifs

1 Efthymios Tzinis and Alexandros Potamianos, “Segment-based speech emotion recognition using recurrent neural networks,”*in Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on. IEEE*, pp. 190–195, 2017.

2 Efthymios Tzinis †, Giorgos Paraskevopoulos †, Christos Baziotis, and Alexandros Potamianos, “Integrating recurrence dynamics for speech emotion recognition,”*in Proceedings of INTERSPEECH (in press)*, 2018.

3 Giorgos Paraskevopoulos †, Efthymios Tzinis †, Emmanuel-Vasileios Vlatakis-Gkaragkounis, and Alexandros Potamianos, “Pattern Search Multidimensional Scaling,”*Under Review for Journal of Machine Learning Research (JMLR), arXiv:1806.00416*, 2018.

† Both authors contributed equally in this work