# Acoustic feature selection for automatic emotion recognition from speech

A sufficient number of training examples is the premise for most machine learning and data mining algorithms to work well. When there are only a few training examples, it is possible to have the problem of overfitting – model can have perfect performance on the training set but cannot generalize well on new examples. If a data set cannot fully cover the whole variable space then it is referred to as a small data set. In this sense the data sets for emotion recognition are small because the typical size is less than 1000 and the number of features is close to 100.

A machine learning framework for emotion recognition

Most existing work can be summarized into the following general procedure:

* Feature extraction stage: extract whole acoustic feature set from the original speech corpus and transform these features into an appropriate format for further processing
* Data preprocessing stage: select the most relevant subset of the whole candidate feature set or reduce the size of the speech data set into fewer dimensions
* Emotion recognition stage: apply machine learning methods on the processed speech data set from the previous stage to recognize emotional states in speech.

# Chapter 22: Learning by training neural networks

*Italics denotes where I have paraphrased and it is not same wording in book*

Each link is associated with a weight which determines the nature and strength of one nodes influence on another. One nodes influence on another is the product of the influencing neurons output value time the connecting links weight. Large positive weight = strong excitation.

Each node combines separate influences received on its input links into an overall influence using an activation function. Simple activation function example is just using a threshold function, output is either 0 or 1 based on if above or below threshold.

## Back propagation does hill climbing by gradient ascent

Adjust weights to improve output of NN – move in direction of most rapid performance improvement by varying all weights simultaneously *in proportion to the improvement observed.* This is gradient ascent. BP computes changes to weights in final layer first then reuses much of the computation to *change weights in preceding layers, sequentially.*

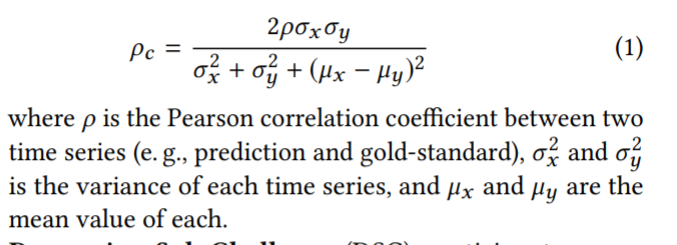
# Automatic Emotion Recognition and classification

Emotions are key semantic component in human communication = effective communication only accomplished when meaning and emotion of communication is understood by all parties involved.

Emotion is a mental state or a feeling that arises subjective.

# AVEC 2017: Real-life Depression and Affect Recognition

Concordance Correlation Coefficient (CCC): evaluates agreement between two time series by scaling their correlation coefficient with their mean square distance.



## Baseline Features

Different acoustic feature sets included, all based on extended version of Geneva Minimalistic Acoustic Parameter Set (eGeMAPS). It is an expert knowledge based feature set consisting of 23 acoustic LLDs extracted over a time frame.

LLDs only capture very local info in time so need segment level representation of features.

Bag of audio words (BoAW) – originally for text. LLDs over a certain segment are first quantised using codebook of “audio words”, then a histogram of audio words occurring in the segment is created.

Video features:

* Face orientation in degrees (3 features, pitch, yaw and roll)
* Pixel co-ords for 10 eye points (for x and y so 20 features in total)
* Pixel co-ords for 49 facial landmarks (x and y so 98 in total)

Segment level representation – bag of video words (BoVW) computed from normalised facial features.

# AVEC 2018: Bipolar Disorder and Cross Cultural Affect Recognition

## Ranking metrics for regression:

CCC is preferred over Pearson’s Correlation Coefficient as it is sensitive to bias and scaling and it permits discriminative training when used as a cost function.

## Facial Recognition idioms and previous results

A common idiom in facial expression recognition is that due in part to consistencies across all humans in terms of make-up of facial muscles [12], emotional expressions have a large degree of universality across cultures. It was hard to find works in affective computing recognition literature which supported this claim [17], until the AVEC 2019 paper lol.

It has been argued that results of facial expression perception studies can be easily biased by the manner in which the answers are elicited [65].

Training with cultures from similar language families has been shown to have better accuracy than dissimilar [28, 67].

## Generation of Emotion Labels

Emotion Recognition requires a large amount of labels of sufficient quality to train systems to learn mapping from input data and labels describing the emotion. Labels need careful attention in definition as they are subjective and highly variable as it depends on human judgement.

Humans seem to be more efficient at discriminating among options that assigning absolute values to subjective variables [50, 84]. The dominant approach in affect modelling relies on absolute values of dimensional attributes such as arousal or valence that are annotated time continuously over the recordings.

Dominant practice – summarise the annotations for each recording into single time series known as the gold-standard. This can easily be processed by any ML algorithm.

Issues during fusion of individual annotations: inconsistencies between values reported by annotators, delay is present between emotional event expressed in the data and the corresponding annotation value.

Method to process noisy time continuous labels reported by humans on dimensional attributes of humans – winning contribution of AVEC 2019 [71], maximise PCC between audio-visual features and the gold standard to estimate the delay used to compensate the **reaction time of annotators.** In AVEC 2015 [61] participants proposed to estimate an overall reaction time for each emotional dimension by maximising recognition performance while varying the delay in a grid search [35, 36].

## Expert knowledge

The traditional approach consists in summarising LLDs of speech and video over time with a set of statistical measures compute over a fixed duration sliding window. Features can be brute forced with large ensemble of LLDs that are all combined with a large set of statistical measures – ComParE acoustic feature set [70], or they can be reduced to smaller expert knowledge based info.

# AVEC 2019: Cross-Cultural Affect Recognition

Dataset: large volume of un-segmented, non-prototypical and non-preselected data of wholly naturalistic behaviour. This is the kind of data that new generation of human-robot communication interfaces have to face in the real world.

Audio-visual recordings collected *“in the wild”*. Standard webcams, at home/work place. Knowledge of German/Hungarian culture leveraged to infer knowledge on Chinese culture.

Ranking on labels relies on CCC – most suitable as it is not biased by changes in scale and location and elegantly includes information on both precision and accuracy in a single statistical measure.

CCC can be easily exploited as a loss function for training neural networks.

## Current State of the Art in cross cultural emotion recognition

A common idiom in facial expression recognition is that emotional expressions have a large degree of universality across cultures.

Vision only or multi modal systems achieved higher cross culture accuracies than speech only approaches.

Techniques used in AVEC 2018:

* Model based on emotional salient detection to identify emotion markers invariant to socio cultural context.
* Data driven approach based on long short term memory recurrent neural networks LSTM-RNN

All entrants in AVEC 2018 observed a drop in system performance when testing on Hungarian data.

## Cross-cultural Emotion Database (SEWA)

Spontaneous behavior

Video chats annotated w.r.t emotional dimensions: *arousal, valence* and *liking* (sentiment). Annotation contours (traces) are combined into a single gold-standard using *evaluator weighted estimator*.

## Baseline Features

Audiovisual representations can be learnt from expert driven information extracted from raw signals or directly from the raw signals.

Also can be generated using adversarial networks or CNNs trained on out-of-domain data and for a different task e.g. audio representations extracted by a model trained for objects classification in images.

Traditional approach in affect sensing consists in summarising low level descriptors of audiovisual signals over time with a set of measures computed over a sliding analysis window. For audio channel descriptors include spectral, cepstral, prosodic and voice quality info. For video channel they are appearance geometric and FAUs.

## Bag-of-Words

Represents distribution of LLDs according to a dictionary learned from them.

<https://machinelearningmastery.com/gentle-introduction-bag-words-model/>

## Deep Representations

Deep spectrum features – audio baseline feature representation.

Inspired by deep representation learning paradigms common in image processing. Spectral images of speech instances are fed into pre-trained image recognition CNNs, resulting activations are extracted as feature vectors

## Cross Cultural Sub challenge

Results of AVEC 2019 baseline features confirmed the idiom that facial expressions of emotion have a large degree of universality across cultures compared to vocal expressions where the acoustic and prosodic dimensions already play a key role in communication e.g. tonal languages like Mandarin.

# Chen, Multimodal Multi-task Learning for Dimensional and Continuous Emotion Recognition

Dimensional emotion theory considers an emotion state as a point in a continuous space described by arousal (measure of affective activation), valence (measure of pleasure), and dominance (measure of control).

These guys found the temporal LSTM model significantly outperformed the non-temporal SVR model for arousal and valence dimensions.

# Chorowski, attention based models for speech recognition

*Has good performance on sequence to sequence network of image captioning.*

Iteratively process their input by selecting relevant content at every step.

Performance degraded quickly with longer concatenated utterances.

**General Framework**

An attention based recurrent sequence generator is a recurrent neural network that stochastically generates an output sequence from an input x. In practice, x is often processed by an encoder which outputs a sequential input representation more suitable for the attention mechanism to work with.

LSTM and GRUs are typically used as a recurrent activation

# Continuous Emotion Recognition: Another Look at the Regression Problem

Regression is an optimization problem that is used to set model parameters so that the resulting model minimizes the prediction error. If the criterion for the goodness of a regression model is other than the prediction error we might need to modify the cost function of the optimization. A commonly used measure for assessing the goodness of a prediction is the correlation coefficient of prediction values with the actual value of a response value.

Perfect prediction = maximum correlation coefficient.

Lower prediction error does not guarantee a higher correlation coefficient.

# Dieterrich, Advances in neural information processing systems 14 vol 2

# Vol 2

Reinforcement learning with Long Short Term Memory, Bram Bakker, Leiden University

LSTM originally designed for supervised timeseries learning. Based on the analysis of the problems that conventional recurrent neural network learning algorithms e.g. BPTT or RTRL have when learning timeseries with long term dependencies – errors propagated back in time tend to vanish or blow up.

LSTMs solution is to enforce constant error flow in a number of specialized units called Constant Error Carrousels (CECs). CECs have linear activation functions that do not decay over time. To prevent CECs from filling up with useless info access is regulated with input gates. Input gates receive input from time series and other units in the network and they learn to open and close access to CECs at appropriate moments. Access from the activations of the CECs to the output units is regulated with output gates which also learn when to send an output. Also have forget gates which reset the activation when the info in CEC is no longer useful. Combination of CEC and gates is a memory cell.

# Dreyfus, Neural Networks Methodologies and applications

Learning for Recurrent Networks

The main problem with RNN learning using a descent method comes from the time range of the consequences of changing a weight value. The influence of a weight value on the cost function is not limited to the current time, it propagates. Training an RNN by propagating the computation for each input, computing the weight correction and iterating would be very expensive for long training sequences such as videos, difficult to implement on real time systems.

Two training methods:

* Compute the true gradient with respect to the current weights but change the cost function by truncating the computation period to a sliding window. This is Back propagation through time.
* Approximate the gradient of previous states with respect to the current weights by the values of those gradients with respect to the previous weights == Real time recurrent learning.

# Hochreiter, LSTM

LSTMs computational complexity per time step and weight is O(1).

Recurrent networks can use their feedback connections to store representations of recent input events in form of activations (“short term memory” as opposed to “long term memory”, embodied by slowly changing weights).

With conventional BPTT or RTRL error signals flowing backwards in time either blow up or vanish.

The temporal evolution of the back propagated error exponentially depends on the size of the weights.

LSTM enforces constant (neither exploding nor vanishing) error flow through internal states of units. Gradient computation is truncated at certain architecture specific points, does not affect long term error flow.

# Attention-Based Multimodal Fusion for Video Description Chiori Hori

Current methods for video description are based on encoder-decoder sentence generation using RNNs.

Temporal attention mechanism gives more weight to encoded features from specific time frames. Use two different types of features : image features and motion features.

**IN this paper they found that combining multimodal (image feature, motion feature and audio feature) attention model with standard temporal attention outperforms state-of-the-art methods on two standard datasets.**

Benefits of attentional multimodal fusion:

* Modalities that are most helpful to discriminate *emotion* can dynamically receive a stronger weight
* Network can detect interference e.g. noise and other sources of uncertainty in each modality and dynamically down-weight the modalities that are less certain.

***GOOD FOR IN THE WILD THEN AMIRITE?***

# Lipton, A critical review of recurrent neural networks for sequence learning

In the setting of large datasets, simple linear models tend to under fit and often underutilize computing resources. Deep learning methods which exploit the local dependency of visual information have demonstrated record setting results on many important applications. (on CNNs)

RNNs are connectionist models with the ability to selectively pass information across sequence steps while processing sequential data one element at a time.

**Why explicitly model sequentiality?**

SVMs, logistic regression and feedforward networks have proved immensely useful without explicitly modelling time. Many models implicitly capture time by concatenating immediate predecessors and successors thus presenting a sliding window.

However not long term memory, keeping context e.g. extended dialog in a call centre automation, remembering complete context of the conversation.

# Basic concepts of Artificial Neural Networks (Winston AI book)

ANN consists of processing elements (units), a connection topology and a learning algorithm. Associated with each input connection link of the unit is a weight value.

Connection topology takes two forms – feed forward only and feedback or recurrent architecture.

# Introduction to neural networks, Phil Picton

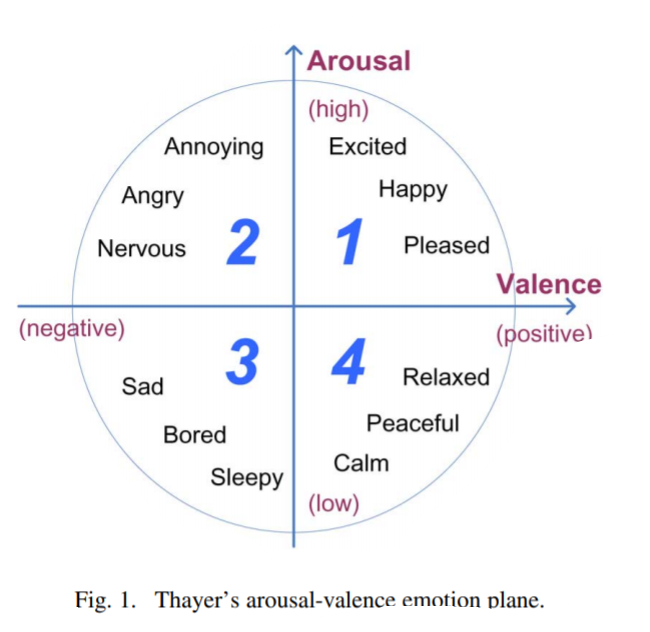
Network of interconnected elements

An RNN is one where feedback is allowed, particular type of feedback network. Inputs arrive at each iteration. Neurons update synchronously (all fire together). Training RNN is more complex, involves many iteration.

Two major training methods are back propagation through time and real time recurrent learning.

# Yang, A regression approach to Music Emotion Recognition

Regression approach is free of the ambiguity inherent to conventional categorical approaches



Not easy to describe emotion in a universal way because adjectives used may be ambiguous and can vary from person to person

**Have a look at support vector regression yeah**

Regression theory => predicting a real value from observed variables/ features