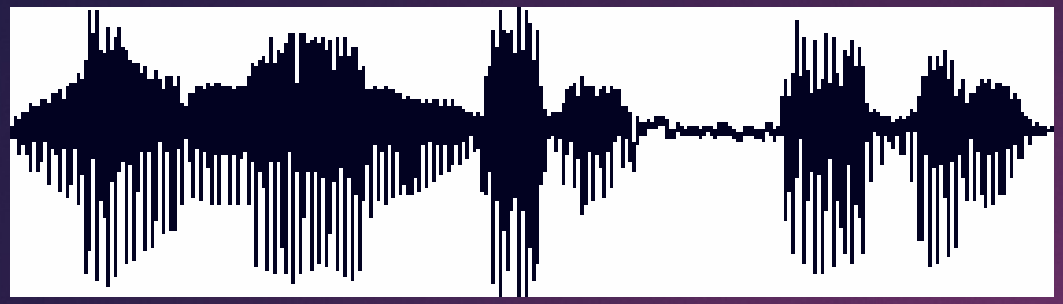


Harmony Search for feature Selection in Speech Emotion Recognition

PANKAJ CHAUHAN (18)

GAURANG DATE (21)

NEHA NAGARKOTI (74)

GUIDED BY: KEVIN D’SOUZA

Problem Statement

- Speech emotion recognition is a particularly valuable for many real time applications. High dimensional data sets create problems even for automated systems.
- In this project, our approach is to select a small subset out of the thousands of speech Data which is important for accurate classification of speech emotion recognition using Harmony Search Algorithm.

FEATURE SELECTION

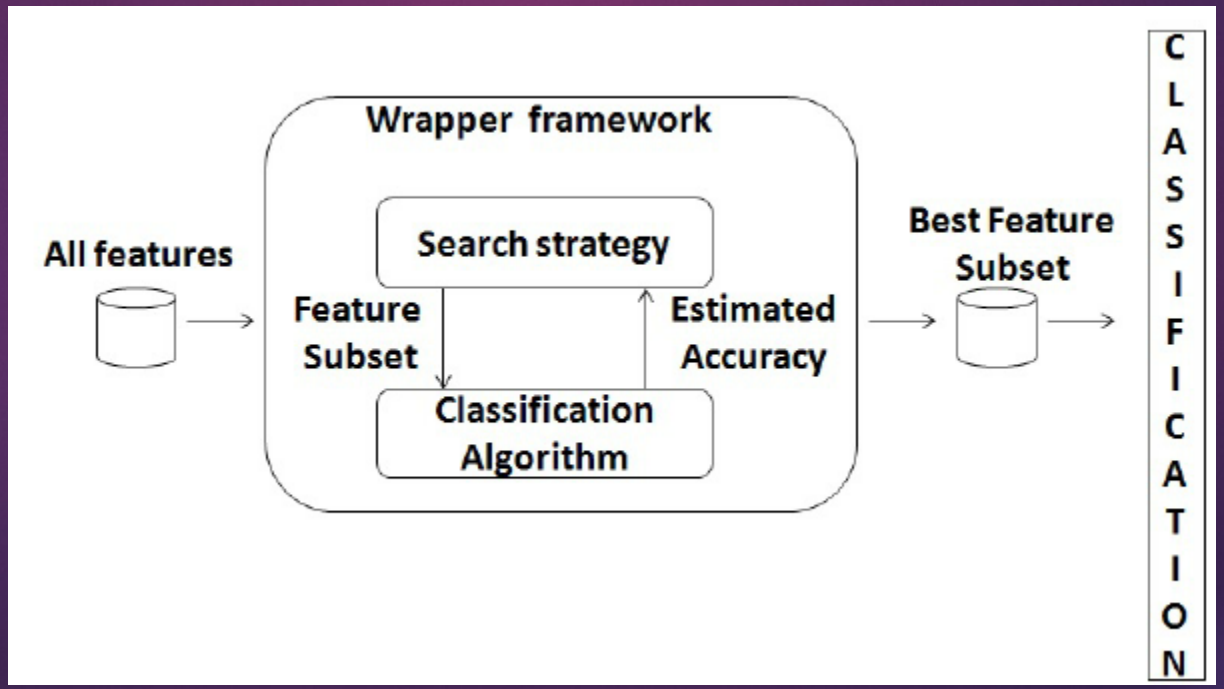
Feature Selection vs Dimensionality Reduction

- In Feature Selection , we simply mute / remove the features irrelevant to us without changing them.
- In Dimensionality Reduction (such as PCA) , the number of features are reduced by making combinations of our existing features .

- There are two main **approaches** for **feature selection** namely:

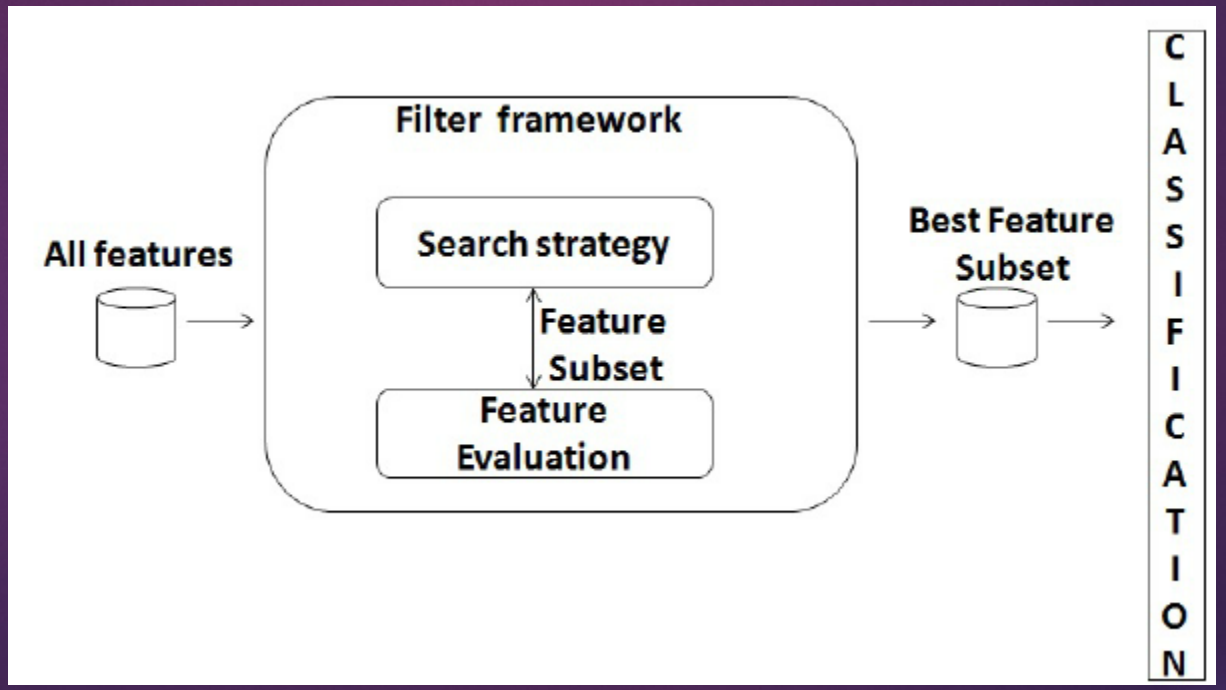
Wrapper method

- It is method in which the **features** are selected using the classifier.
- It is mainly used as post process method. ie works on features after classification is done to optimize it.



Filter method

- It is method in which the **selection** of **features** is independent of the classifier used.
- They are mainly used as a pre-process method.



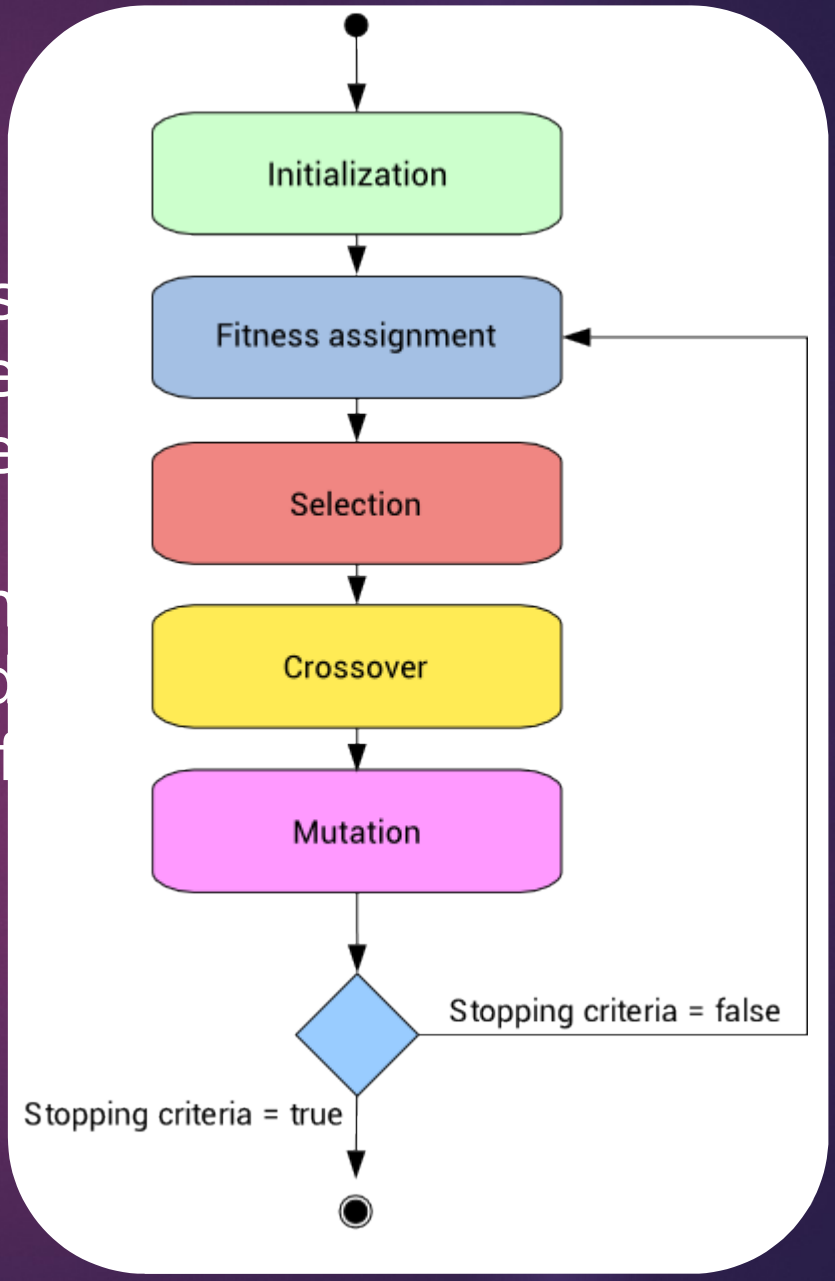
A Survey on Evolutionary Computation Approaches to Feature Selection

Genetic Programming

- GP is used more often in feature construction than feature selection because of its flexible representation.
- GP can be used as a search algorithm and also a classification algorithm.

Genetic Algorithm

- In nature, the genes of organisms tend to evolve over successive generations to better adapt to the environment.
- The Genetic Algorithm is an heuristic optimization method inspired by that procedures of natural evolution.



COMPARISION

- For Bangla text recognition SVM classifiers were used and to optimize Fuzzy feature data set HS, GA and PSO algorithm were tested.
- Having an average accuracy of 89% with all features and higher computation time was reduced drastically.
- Also improvisation in accuracy was observed with reduced computation time.

Optimization Algorithm	Optimization Feature Subset	Classification Accuracy (%)	Execution time (sec)
Genetic Algorithm	45	84.65	1509.25
Particle Swarm Optimization	40	85.19	1248.89
Harmony Search Algorithm	48	90.29	944.75

Advantage of HS

Over Genetic Algorithm

- Harmony Search algorithm is more fast.
- Randomization is considered in Harmony search.
- Better accuracies are provided for same table.
- HS optimization is more easier than GA

Over Particle Swarm Optimization

- Individual features are considered.
- This helps to give better accuracy than POS.

WHY HARMONIC SEARCH

- Existing Nature inspired Algorithms:
  - Genetic Algorithm (GA)
  - Evolutionary Algorithm (Evolution)
  - Simulated Annealing (Metal Annealing)
  - Ant Algorithm (Ant’s Behavior)

- Diversification
- Intensification
- relatively easier
- versatile to combine HS with other metaheuristic algorithms
- Parallelism possible Higher efficiency

HARMONIC SEARCH

- Harmony search is a music-based metaheuristic optimization algorithm. It was inspired by the observation that the aim of music is to search for a perfect state of harmony.

- When a musician is improvising, he or she has three possible choices:
  - (1) playing any famous tune exactly from his or her memory

USAGE OF HARMONY MEMORY

  - (2) playing something similar to the aforementioned tune (thus adjusting the pitch slightly)

PITCH ADJUSTING

  - (3) composing new or random notes

RANDOMIZATION

Harmonic Memory

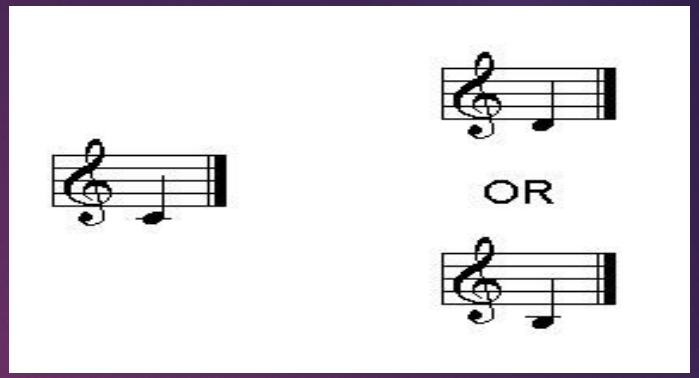
- It ensures that good harmonies are considered as elements of new solution vectors.
- $r_{accept} \in [0,1]$

  harmony memory considering rate (HMCR)
- $r_{accept} = 0.7 \sim 0.95$
- WHY??

rate low only few elite harmonies are selected and it may converge too slowly

rate extremely high (near 1) Pitches in the Harmony memory are mostly used other ones are not explored well

Pitch Adjusting



- changing the frequency, it means generating a slightly different value in the HS algorithm.

$$x_{new} = x_{old} + \varepsilon \times b_{range}$$

- This action produces a new pitch by adding small random amount to the existing pitch .
- ε is a random number from uniform distribution with the range of [-1, 1].
- PAR =Pitch Adjusting rate

Randomization

- To increase the diversity of the solutions.
- Although the pitch adjustment has a similar role, it is limited to certain area and thus corresponds to a local search.
- The use of randomization can drive the system further to explore various diverse solutions so as to attain the global optimality.

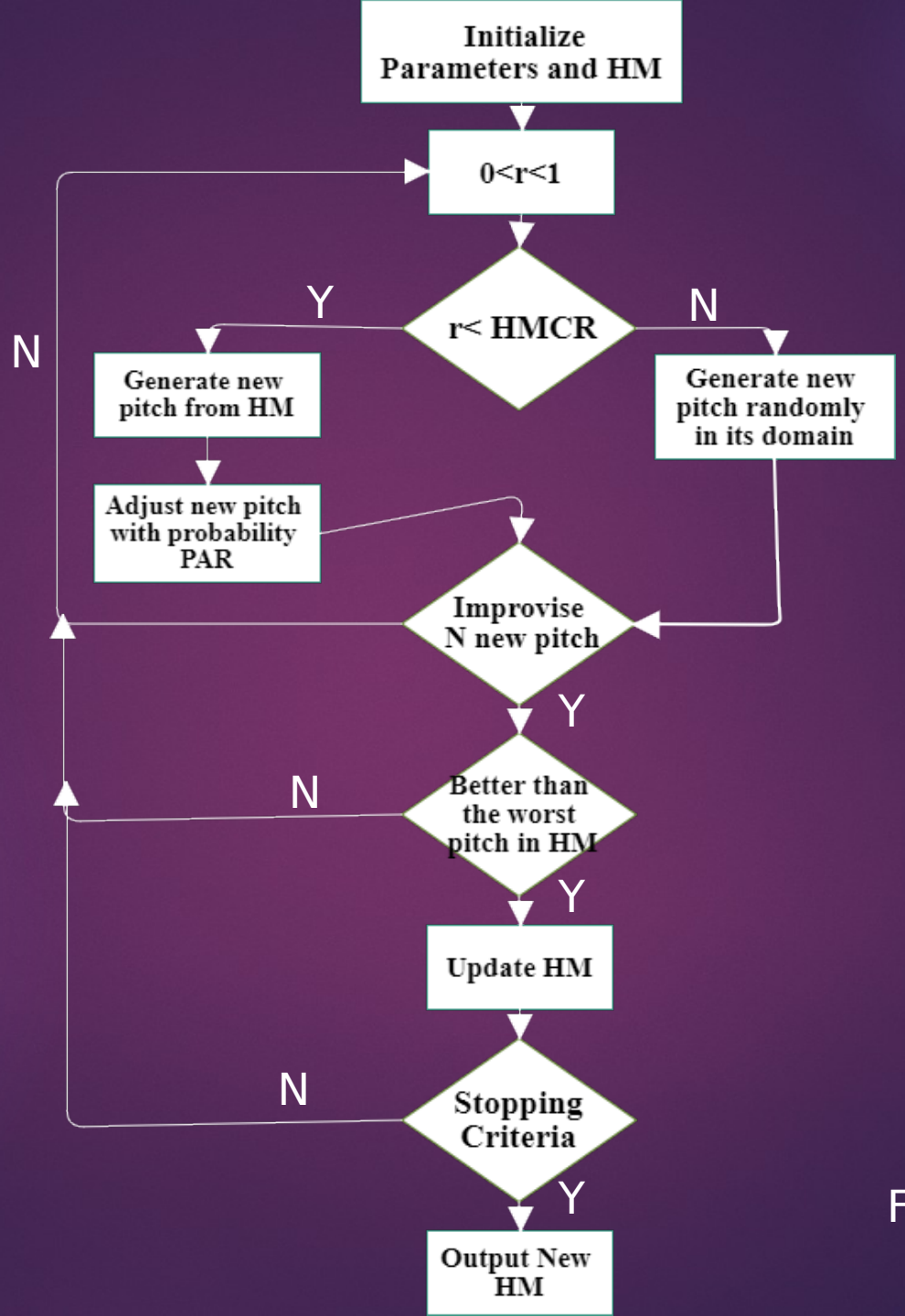


Fig.1: Flow Chart Of Harmony Search Alogorithm

Example:

$$\text{Minimise } (a-2)^2 + (b-3)^4 + (c-1)^2 + 3$$

where $a, b, c \in \{1, 2, 3, 4, 5\}$

Initialize Parameters: HMCR, PAR, Max. Iteration, HM size

No of musicians = No of Variables=3 $\{ p_1, p_2, p_3 \}$

HM Size = 3

Initialize Harmony Memory Randomly

Tone Domain

Select Randomly form domain

Worst Harmony

New Harmony

p_1	p_2	p_3	F
2	2	1	4
1	3	4	8
5	3	3	16
1	2	3	9

Simulation of Optimization problem

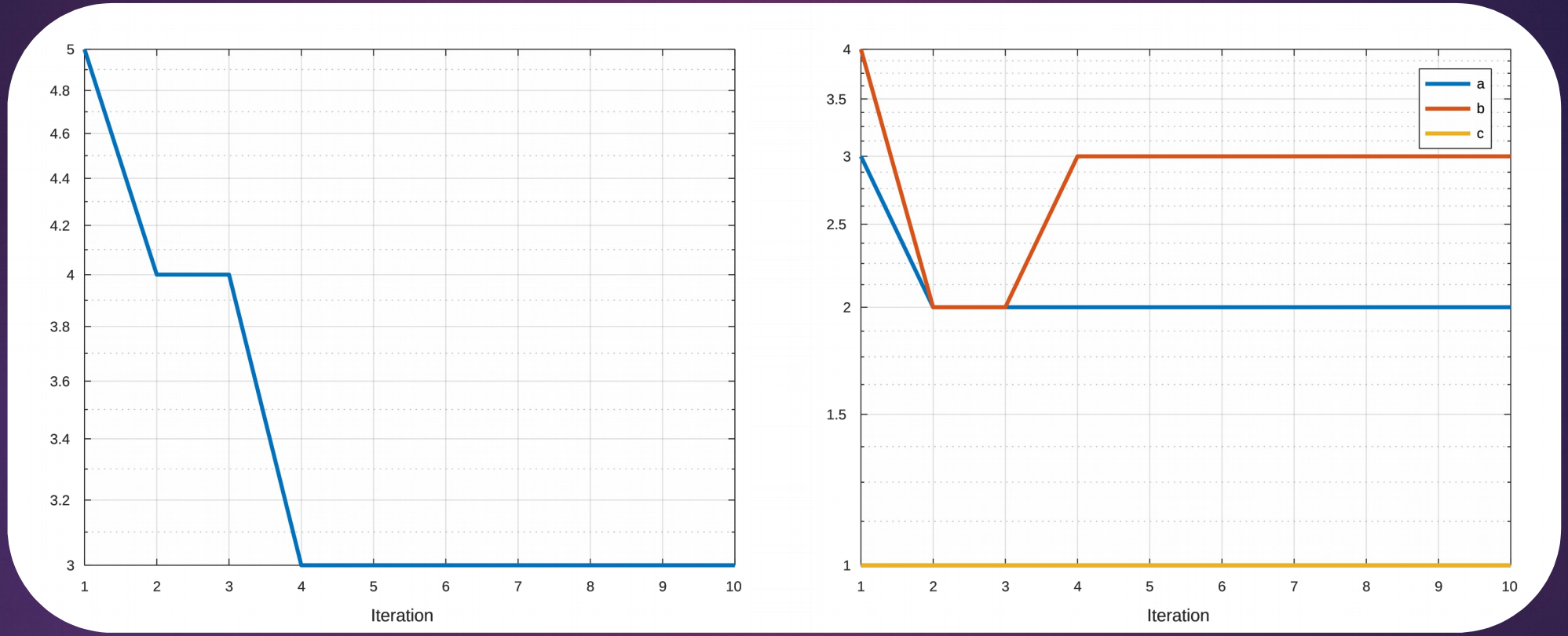


Fig. 1

Mapping from HS to FS

HS	Optimisation	FS
Musician	Variable	Feature Selector
Musical Note	Variable Value	Feature
Harmony	Solution Vector	Subset
Harmony Memory	Solution Storage	Subset Storage
Harmony Evaluation	Fitness Function	Subset Evaluation
Optimal Harmony	Optimal Solution	Optimal Subset

Binary-Valued Representation

- In this method ,we initialize HM randomly with 0 and 1. It represents selecting corresponding position's value for number 1, 0 denotes no selection.
- No of Musicians = No. of features
- Tone domain for each musician will be same i.e. { 0,1 }.
- Suppose feature dataset { f_1, f_2, f_3, f_4, f_5, f_6 }

	p^1	p^2	p^3	p^4	p^5	p^6	Represented Subset
H^1	0	1	0	1	1	0	{f_2, f_4, f_5}
H^2	1	1	0	0	1	1	{f_1, f_2, f_5, f_6 }

. . .

. . .

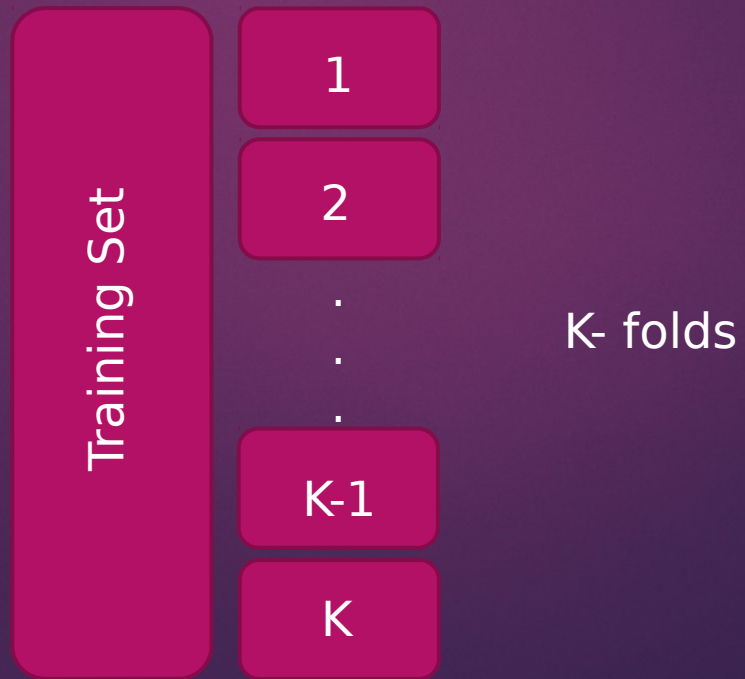
Pitch adjustment and Randomization in Binary-valued representation

- In this method, Pitch adjustment can be done by simply complementing that bit.
- In Randomization , Musician select randomly from { 0,1 }.

	p^1	p^2	p^3	p^4	p^5	p^6	Represented Subset
H^1	0	1	0	1	1	0	$\{f_2, f_4, f_5\}$
H^2	1	1	0	0□1	1	1	$\{f_1, f_2, f_4, f_5, f_6\}$

Cross Validation

- The **purpose** of **using cross-validation** is to make one more confident to the model trained on the training set.
- Cross-validation is a technique to evaluate predictive models by partitioning the original sample into a training set to train the model, and a test set to evaluate it.



Speech Emotion Recognition System

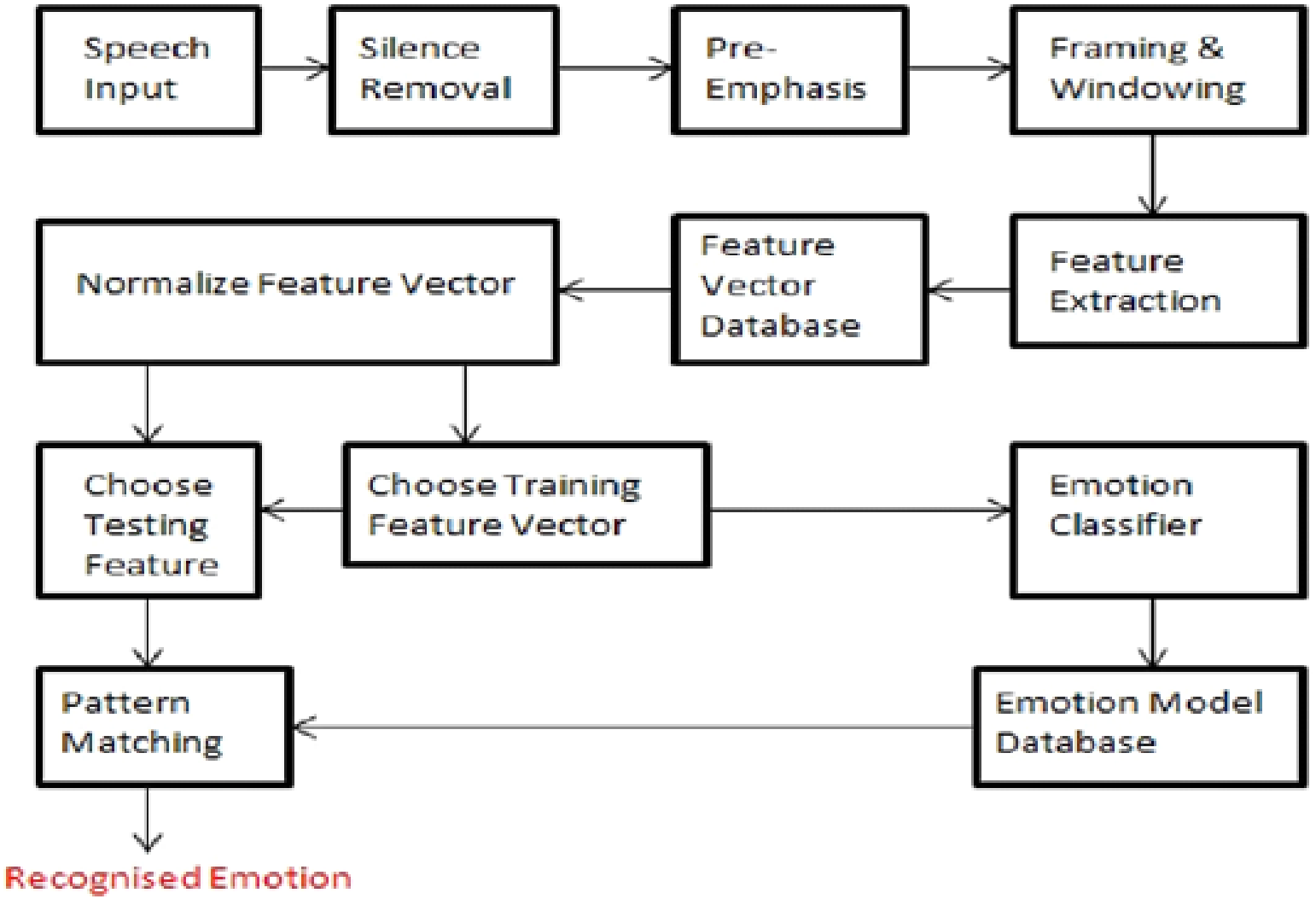


Fig. 2: Block diagram of Speech recognition

Features

MFCC

- Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC.
- The mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.

Pitch

- Pitch is the perceived fundamental frequency of a speech signal, pitch is dependent on the tension created on the vocal fields due to the non-linear air flow during speech generation. It can be extracted using autocorrelation, cepstrum and the most popular RAPT (Robust Algorithm for Pitch Tracking).

Fourier parameters

- Fourier series is one of the most principal analytical methods and has been extensively applied for signal processing, including filtering, correlation, coding, synthesis and feature extraction for pattern identification.
- In Fourier analysis, a signal is decomposed into its constituent sinusoidal vibrations. A periodic signal can be described in terms of a series of harmonically related (i.e., integer multiples of a fundamental frequency) sine and cosine waves.
- The harmonic part of the model is a Fourier serial representation of a speech signal’s periodic components. When a non periodic component is sampled, its Fourier transform becomes a periodic and continuous function of frequency.

Extraction of MFCC Features

- The **Mel-frequency cepstrum** (**MFC**) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a non linear mel scale of frequency.

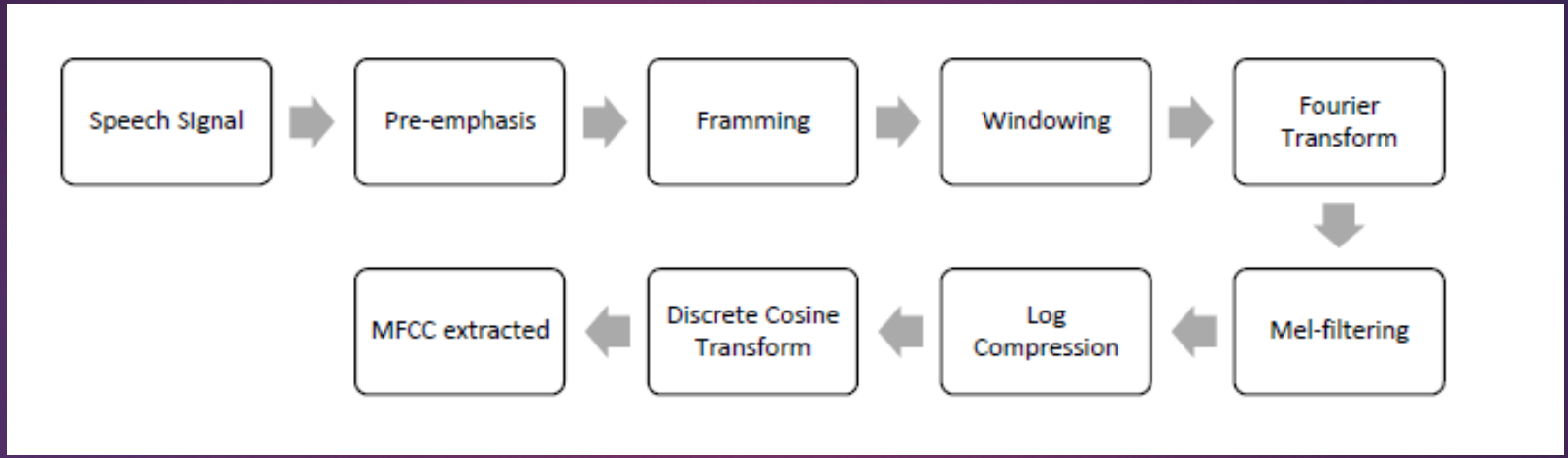
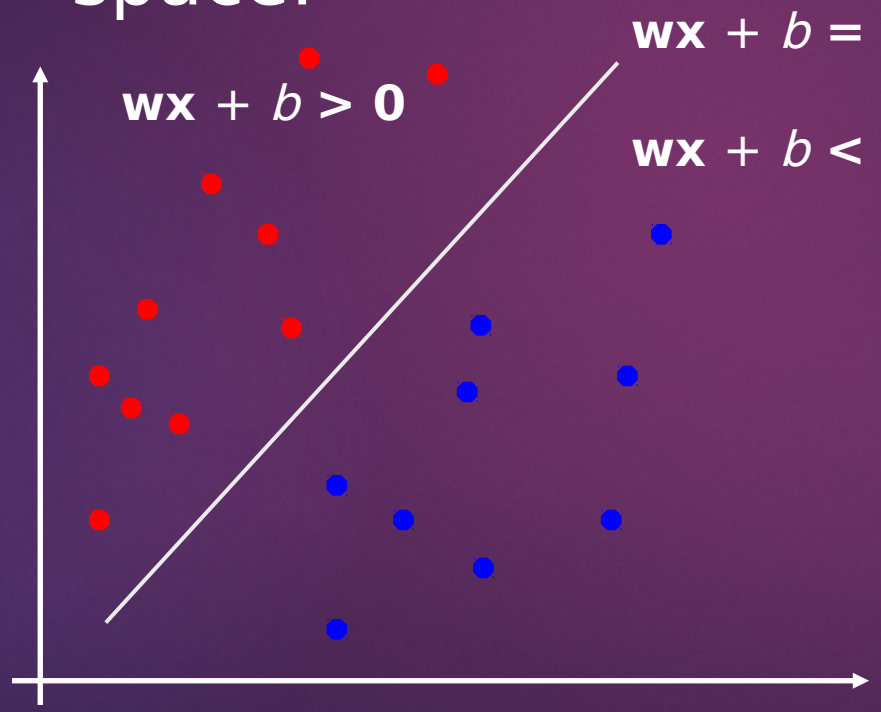


Fig. 3: Block diagram

FEATURE CLASSIFICATION

Linear Separators

- Binary classification can be viewed as the task of separating classes in feature space:



$f(\mathbf{x}) = \text{sign}(\mathbf{w}\mathbf{x} + b)$

Fig. 4

Linear Separators

- Which of the linear separators is optimal?

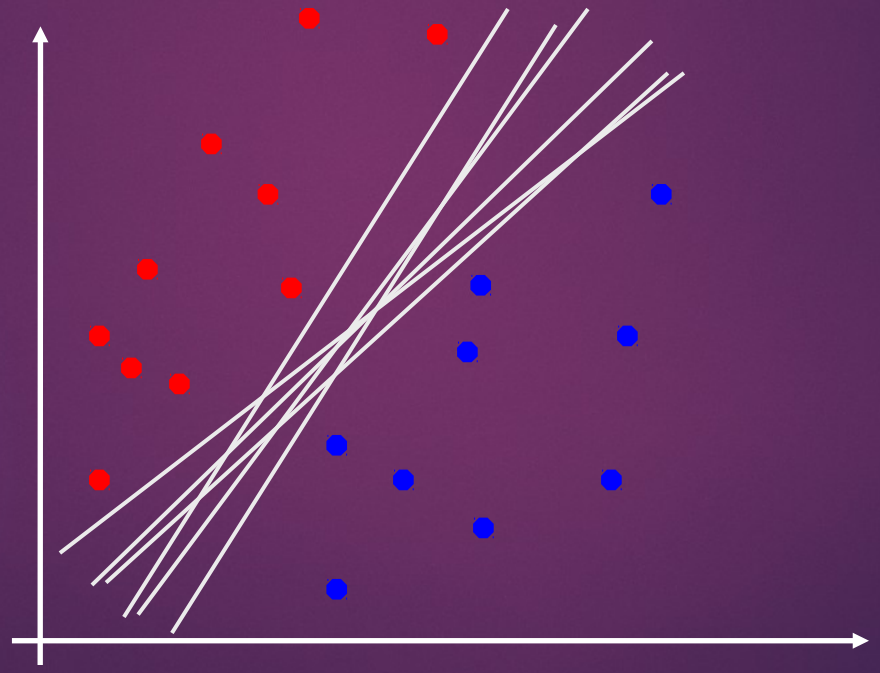


Fig. 5

What is SVM?

- It is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems.

What does it do?

- In this algorithm, we plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate.
- Then, we perform classification by finding the hyper-plane that differentiate the two classes very well.

Why SVM?

- The benefit is that one can capture much more complex relationships between your data-points without having to perform difficult transformations on our own.

Binary SVM: Classification Margin

- Distance from example $\mathbf{x}_i$ to the separator is $r = \frac{\mathbf{w}^T \mathbf{x}_i + b}{\|\mathbf{w}\|}$
- Examples closest to the hyperplane are ***support vectors***.
- ***Margin*** ρ of the separator is the distance between support vectors.

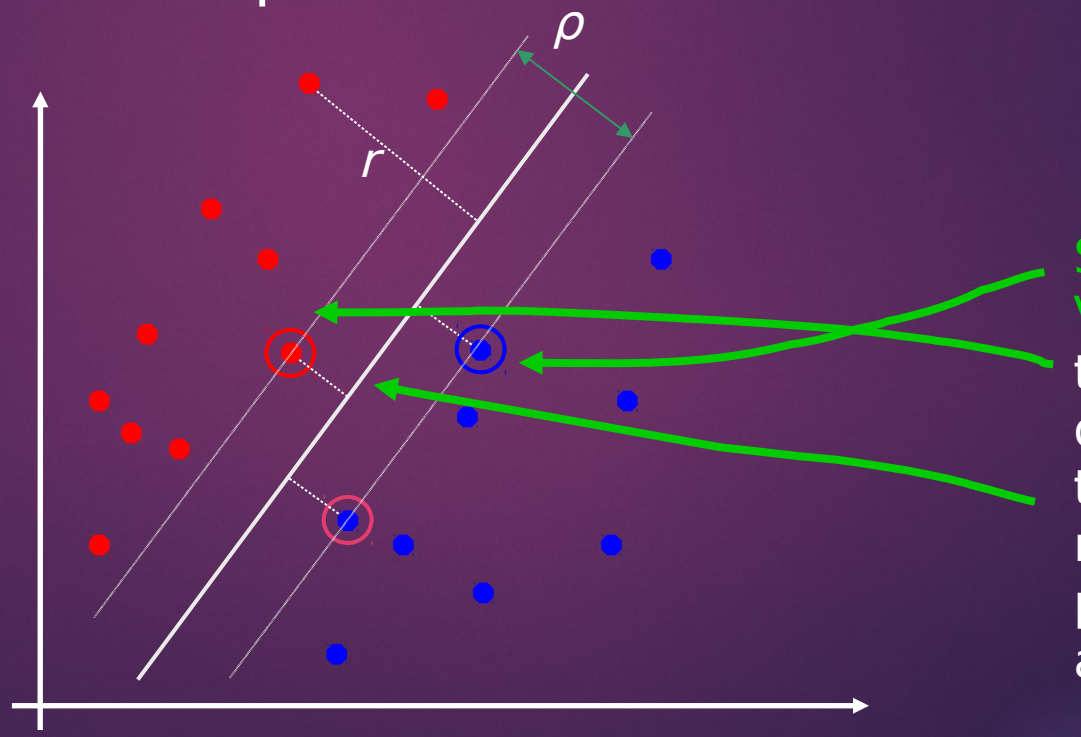


Fig. 6.1

Support Vectors are those datapoints that the margin pushes up against

Maximum Margin Classification

- Maximizing the margin is good according to intuition and PAC theory.
- Implies that only support vectors matter; other training examples are ignorable.

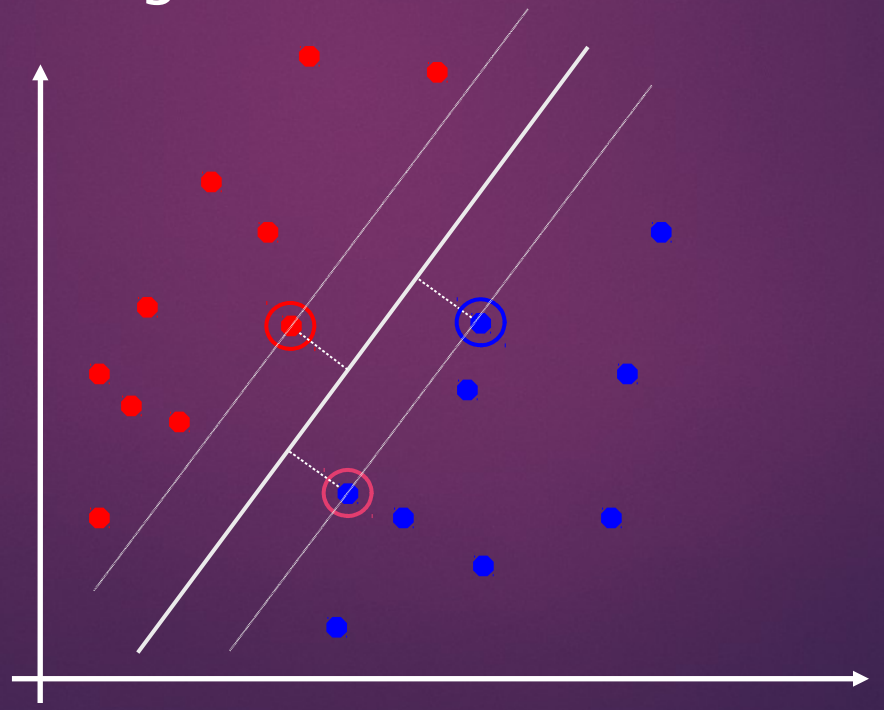


Fig. 6.2

Linear SVM Mathematically

- Let training set $\{(\mathbf{x}_i, y_i)\}_{i=1..n}$, $\mathbf{x}_i \in \mathbf{R}^d$, $y_i \in \{-1, 1\}$ be separated by a hyperplane with margin ρ. Then for each training example $(\mathbf{x}_i, y_i)$:

$$\begin{aligned} &\mathbf{w}^T\mathbf{x}_i + b \le -\rho/2 \quad \text{if } y_i = -1 \\ &\mathbf{w}^T\mathbf{x}_i + b \ge \rho/2 \quad \text{if } y_i = 1 \end{aligned} \Leftrightarrow \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \ge \rho/2$$

- For every support vector $\mathbf{x}_s$ the above inequality is an equality. After rescaling $\mathbf{w}$ and b by $\rho/2$ in the equality, we obtain that distance between each $\mathbf{x}_s$ and the hyperplane is $r = \frac{y_s(\mathbf{w}^T\mathbf{x}_s + b)}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}$

- Then the margin can be expressed through (rescaled) $\mathbf{w}$ and b as: $\rho = 2r = \frac{2}{\|\mathbf{w}\|}$

Problems with linear SVM

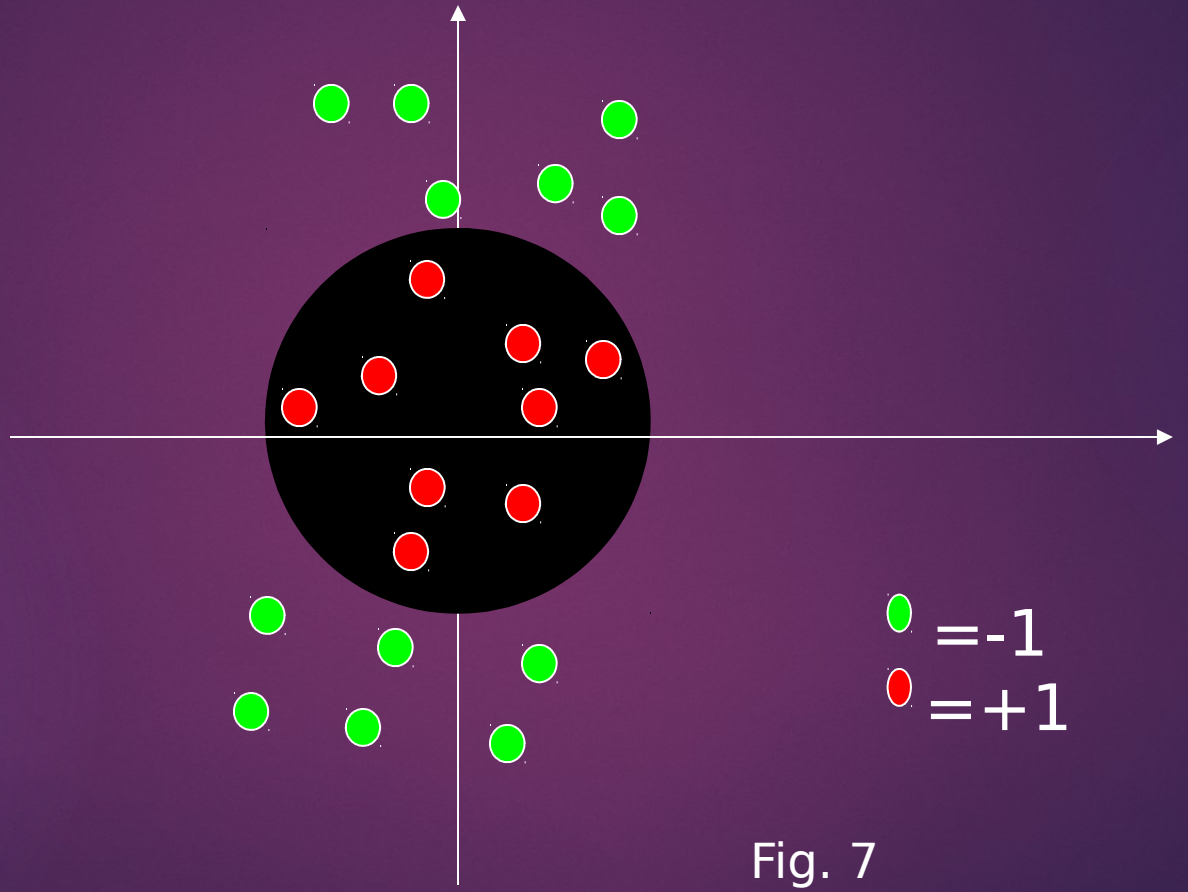


Fig. 7

What if the decision function is nonlinear?

Non-linear SVMs

- General idea: the original feature space can always be mapped to some higher-dimensional feature space where the training set is separable:

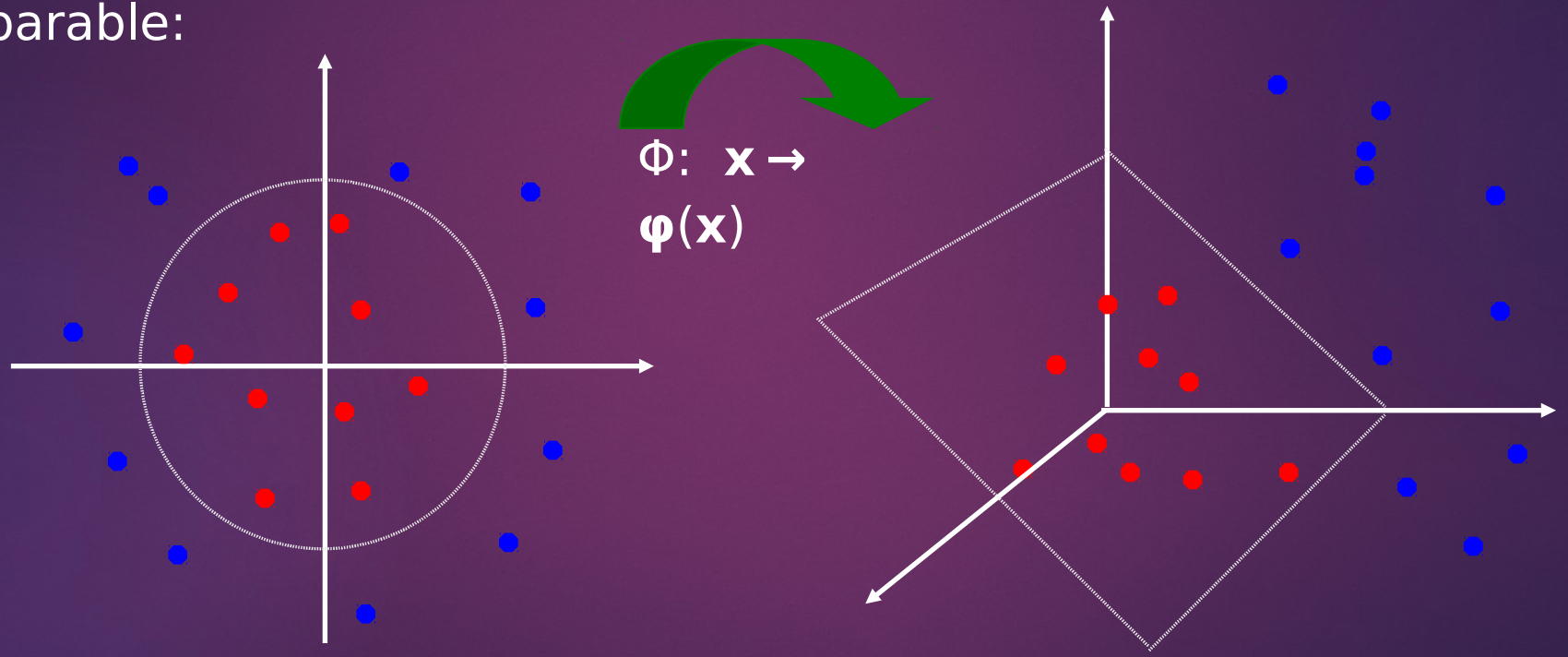
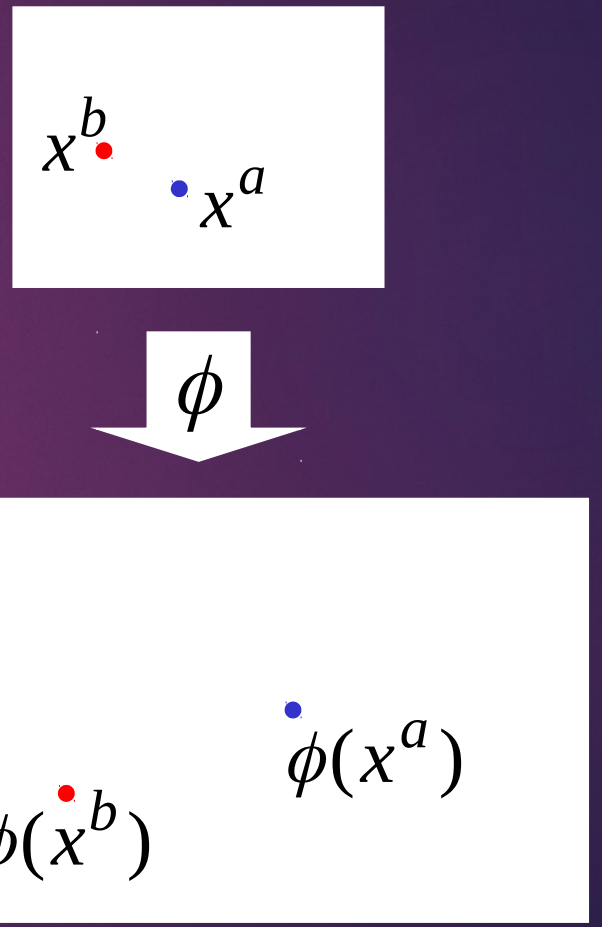


Fig. 8

The kernel trick

- For many mappings from a low-D space to a high-D space, there is a simple operation on two vectors in the low-D space that can be used to compute the scalar product of their two images in the high-D space.



$$K(x^a, x^b) = \phi(x^a).\phi(x^b)$$

Letting the kernel do the work

doing the scalar product in the obvious way

Some commonly used kernels

Polynomial: $K(\mathbf{x},\mathbf{y}) = (\mathbf{x}.\mathbf{y} + 1)^p$

Gaussian radial basis function: $K(\mathbf{x},\mathbf{y}) = e^{-\|\mathbf{x}-\mathbf{y}\|^2/2\sigma^2}$

Neural net: $K(\mathbf{x},\mathbf{y}) = \tanh(k\,\mathbf{x}.\mathbf{y} - \delta)$

Parameters that the user must choose

Doing multi-class classification

- SVMs can only handle two-class outputs (i.e. a categorical output variable with arity 2).
- How to handle multiple classes
- Answer: one-vs-all, learn N SVM's
  - SVM 1 learns "Output==1" vs "Output != 1"
  - SVM 2 learns "Output==2" vs "Output != 2"

    :
  - SVM N learns "Output==N" vs "Output != N"

One-vs-All

- vs the other classes: angry(S)

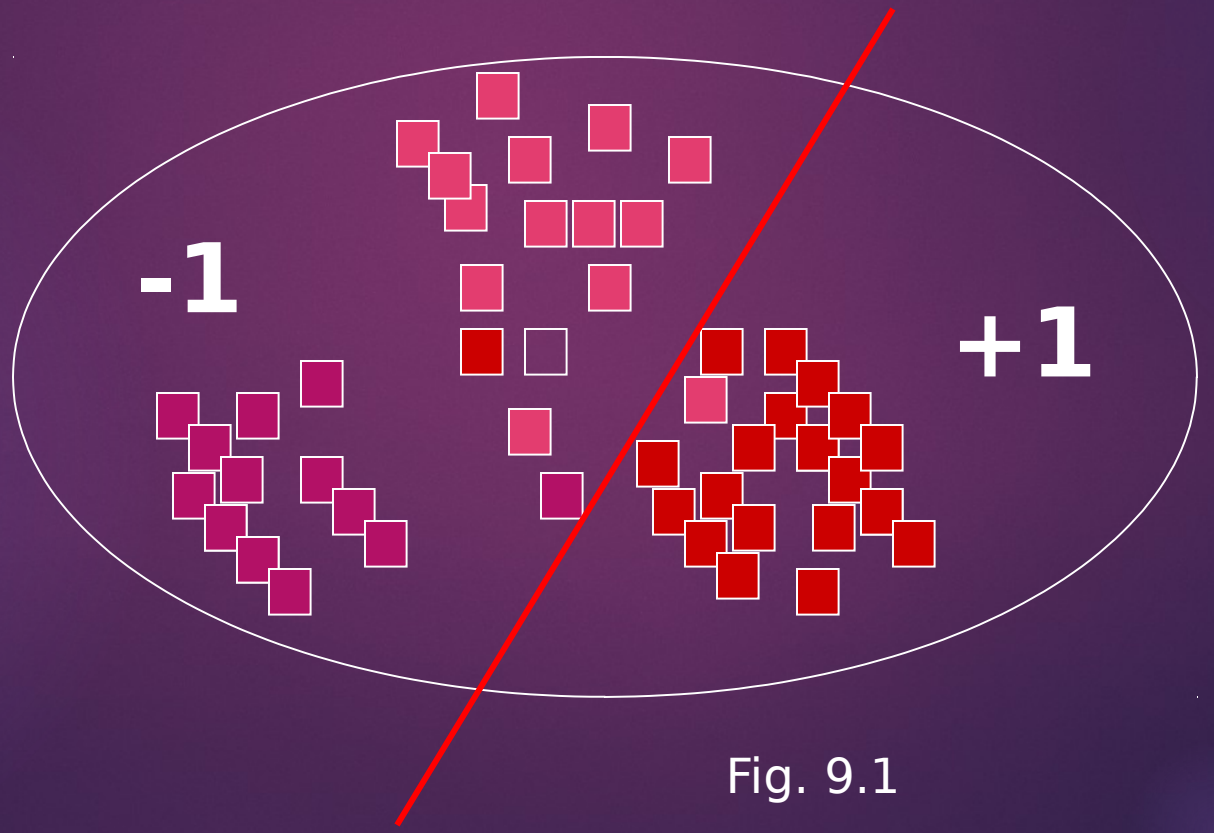


Fig. 9.1

One-vs-All

- vs the other classes: angry(S)
- vs the other classes: sad(S)

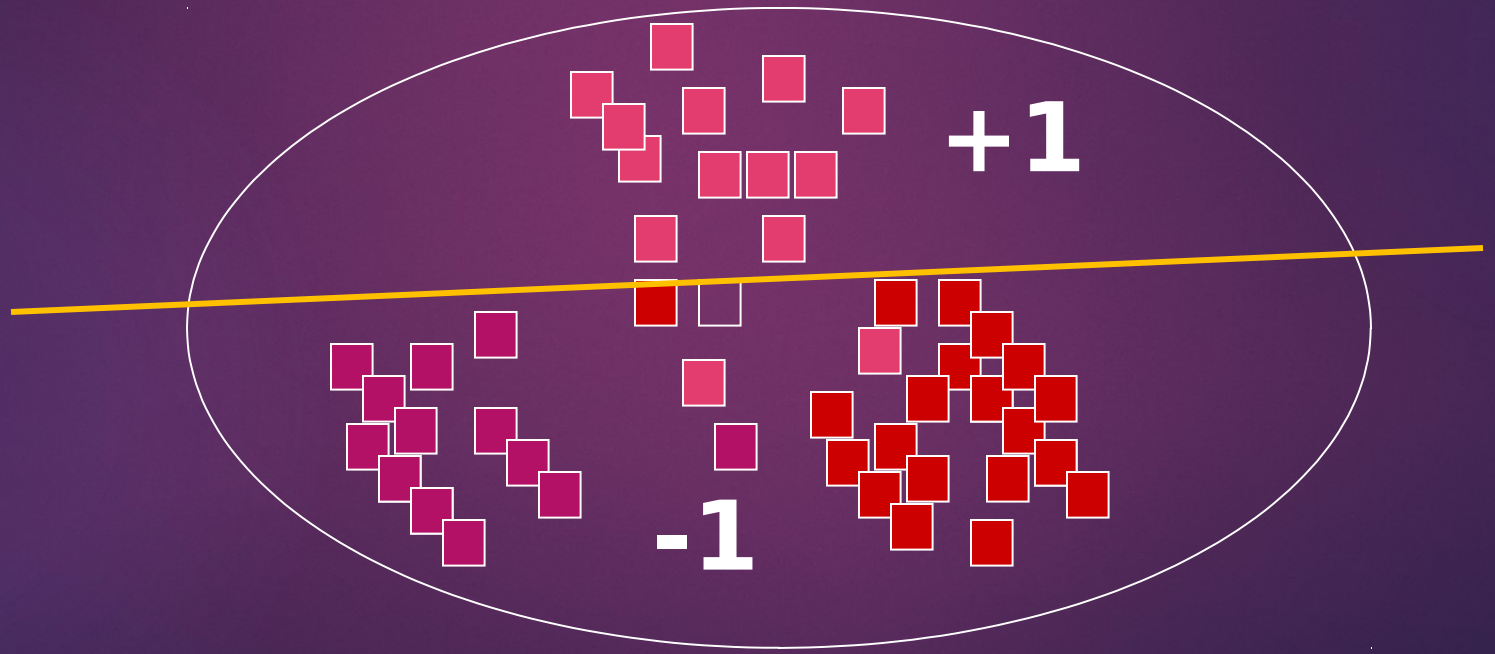


Fig. 9.2

One-vs-All

- vs the other classes: angry(S)
- vs the other classes: sad(S)
- vs the other classes: happy(S)

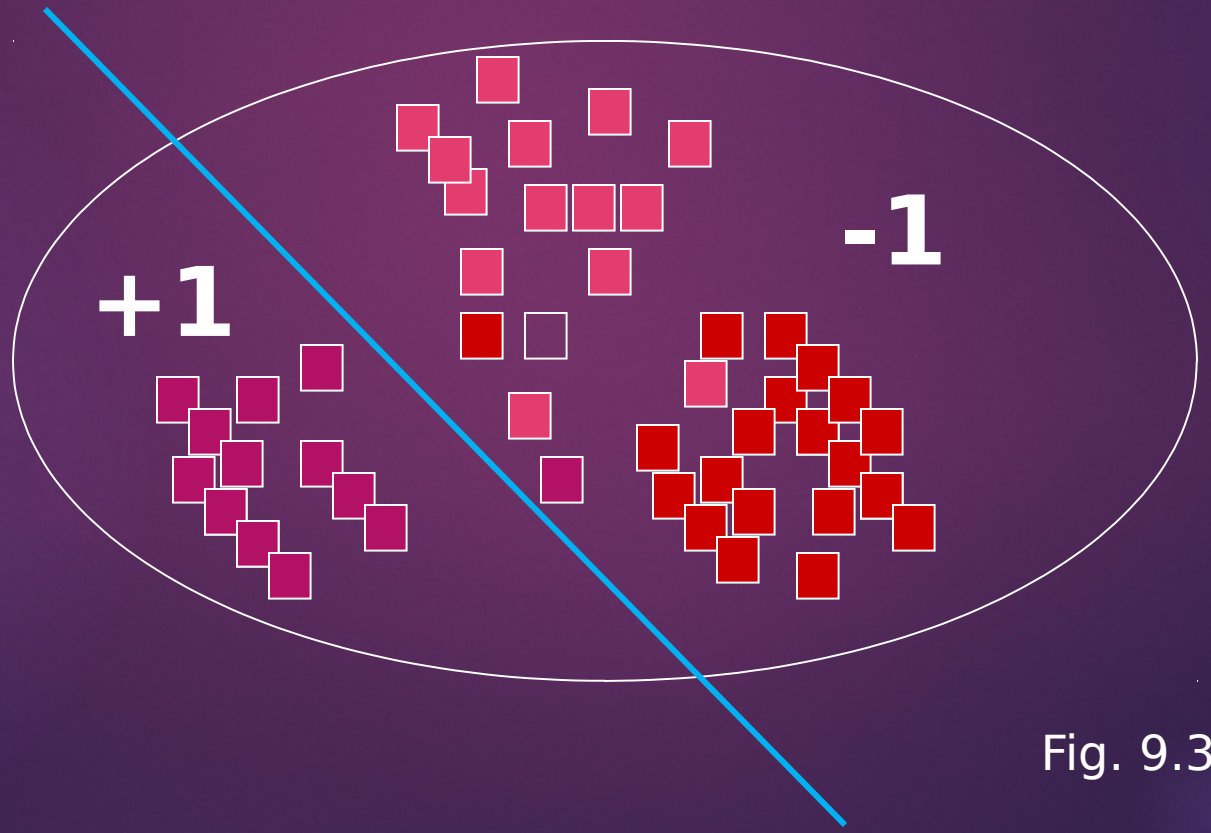


Fig. 9.3

One-vs-All

- vs the other classes: angry(S)
- vs the other classes: sad(S)
- vs the other classes: happy(S)
- Given a test sound S; how to decide its emotion ?
- Assign S to the emotion function with the largest score

Databases

Language	Name of the database	Type of database	Characteristics
German	Emo-DB	Acted	1. 535 Utterances of 5 female and 5 male speakers 2. The age of the speakers range from 21-35 years 3. Anger, boredom, disgust, fear, happiness, sadness and neutral explored.
English	SAVEE	Acted	1. 480 Utterances of 4 male actors 2. The age of the speaker range from 27-31 3. Anger, Disgust, Fear, Happiness, sadness, Surprise, Neutral.
English	SUSAS	Acted	1. 6,000 utterances, 32 actors (13 females + 19 males) 2. Four stress styles: Simulated Stress, Calibrated Workload Tracking Task, Acquisition and Compensatory Tracking Task, Amusement Park Roller-Coaster,

Database Description

	Anger	Boredom	Disgust	Fear	Happiness	Sadness	Neural	Total
EMO DB								
	127	81	46	69	71	62	79	535

Table 1

Experimental Results

Confusion Matrix of Emotion Recognition

SVM (Poly)	OneVsAll		MFCC (13)	(Mean, Std.Deviation, Variance, Kurtosis)			
	'Fear'	'Disgust'	'Neutral'	'Anger'	'Happiness'	'Boredom'	'Sadness'
'Fear'	40.70%	3.70%	3.70%	25.92%	11.11%	14.81%	0
'Disgust'	16.67%	8.33%	33.33%	4.17%	20.83%	16.67%	0
'Neutral'	0	4.31%	56.52%	0	0	34.72%	4.35%
Anger'	0	2.32%	0	93.02%	4.64%	0	0
'Happiness'	0	4%	4%	48%	44%	0	0
'Boredom'	9.67%	6.45%	35.48%	0	0	38.70%	9.67%
'Sadness'	30.43%	4.34%	8.69%	0	0	13.04%	43.47%

Table 2.1

SVM (Poly)	OneVsAll	MFCC (39)(Δ +ΔΔ)		(Mean, Std.Deviation, Variance, Kurtosis)			
	'Fear'	'Disgust'	'Neutral'	'Anger'	'Happiness'	'Boredom'	'Sadness'
'Fear'	18.51%	7.40%	11.11%	55.56%	7.40%	0	0
'Disgust'	0	54.17%	8.33%	4.17%	0	33.33%	0
'Neutral'	0	0	82.60%	4.34%	0	8.69%	4.34%
Anger'	0	0	0	90.69%	9.30%	0	0
'Happiness'	0	0	0	68%	32%	0	0
'Boredom'	0	12.90%	9.67%	0	0	61.29%	16.12%
'Sadness'	0	8.69%	0	0	0	8.69%	82.60%

Table 2.3

Accuracy of Subset over 100 improvisations on EMODB

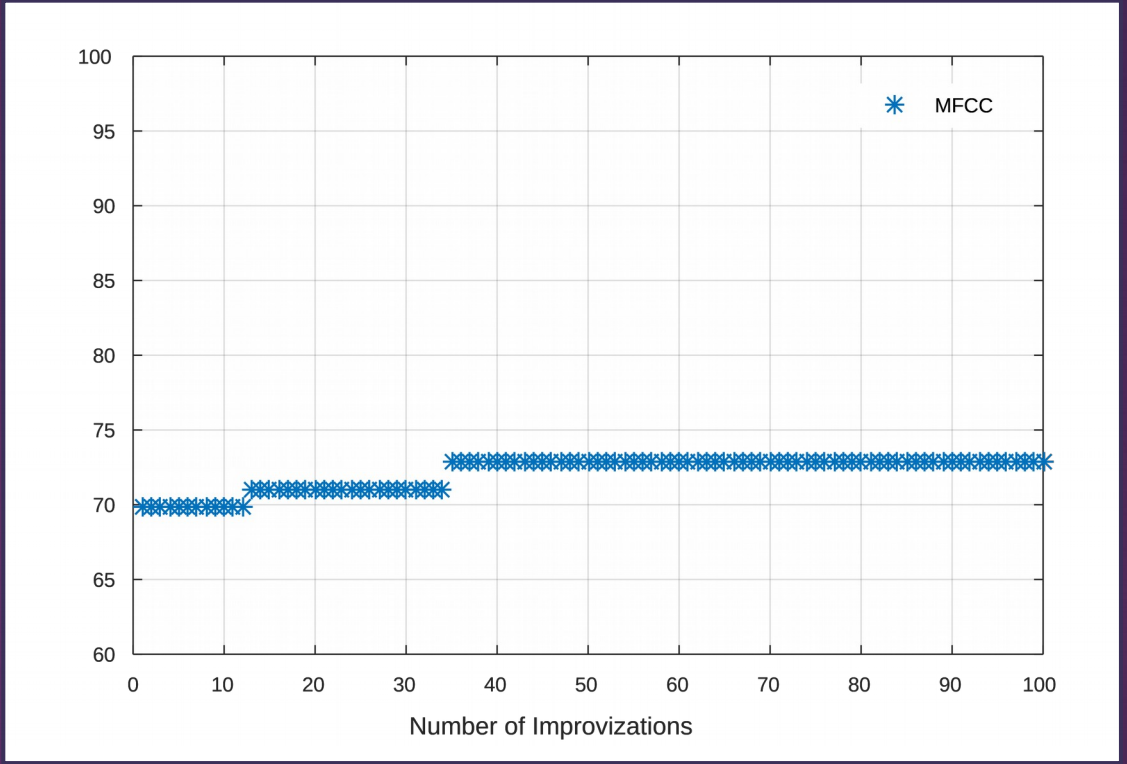


Fig 10

Comparison of selected and unselected feature size on EMODB

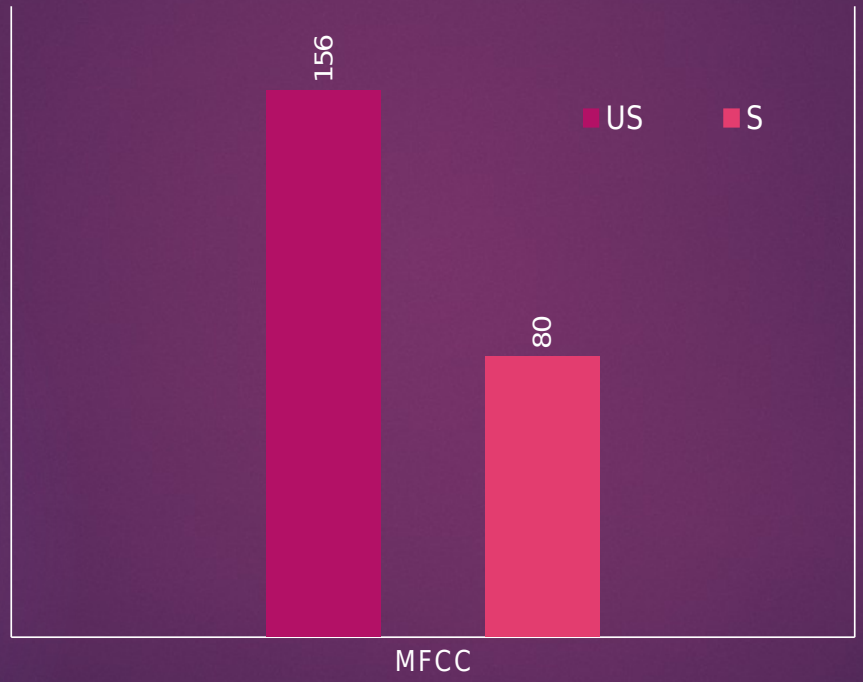


Fig. 11

Comparison of Accuracy‘s and Dimension’s Variation on EMODB

	EMODB					
	Accuracy(%)			Dimension		
	US	S	V	US	S	R(%)
MFCC	70.5	72.86	+2.36	156	80	48.7

WORK FOR SEM 8

Disadvantage of Binary-value Representation

- no. of musicians = total features
  - Scaling problem with large numbers of features.
- Note domain gives each musician very limited choices when composing new harmonies.
- The pitch adjustment opportunities are also wasted for binary choices.
- Takes more time to convergence

- The entire pool of the original features *A*, forms the range of musical notes available to each of the musicians.
- Multiple musicians are allowed to choose the same feature, and they may opt to choose none at all.

	p^1	p^2	p^3	p^4	p^5	p^6	Represented Subset
H^1	f_2	f_1	f_3	f_4	f_7	f_{10}	$\{f_1, f_2, f_3, f_4, f_7, f_{10}\}$
H^2	f_2	f_2	f_2	f_3	f_{13}	-	$\{f_2, f_3, f_{13}\}$
H^3	f_2	-	f_2	$f_3 \square f_6$	f_{13}	f_4	$\{f_2, f_4, f_6, f_{13}\}$

- To eliminate the drawbacks associated with the use of fixed parameter values, a dynamic parameter adjustment scheme is proposed in following paper.
  - R. Diao and Q. Shen, “Deterministic parameter control in harmony search,” in *Proceedings of the 10th UK Workshop on Computational Intelligence*, 2010.
  - So in next Semester , we will use improved Harmony Search method for feature selection.
- Choosing no of musicians become crucial decision.
  - So in next semester, We will try to automate this process.

Refrences

[1] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." Pattern Recognition 44.3 (2011): 572-587.

[2] Tao, Yongsen, et al. "Harmony search for feature selection in speech emotion recognition." Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on. IEEE, 2015.

[3] Gupta, S., Jaafar, J., Ahmad, W.F.W. and Bansal, A., 2013. Feature extraction using MFCC. Signal & Image Processing, 4(4), p.101

[4] Hsu, C.W., Chang, C.C. and Lin, C.J., 2003. A practical guide to support vector classification.

[5] Geem, Z.W. ed., 2009. Music-inspired harmony search algorithm: theory and applications (Vol. 191). Springer.

[6] Abdel-Raouf, O. and Metwally, M.A.B., 2013. A survey of harmony search algorithm. International Journal of Computer Applications, 70(28).

[7] Manjarres, D., Landa-Torres, I., Gil-Lopez, S., Del Ser, J., Bilbao, M.N., Salcedo-Sanz, S. and Geem, Z.W., 2013. A survey on applications of the harmony search algorithm. Engineering Applications of Artificial Intelligence, 26(8), pp.1818-1831.

[8] V. Radisic, Y. Qian, and T. Itoh, “Novel architectures for high-efficiency amplifiers for wireless applications,” IEEE Transactions on Microwave Theory and Techniques, vol. 46, no. 11, pp. 1901–1909, Nov. 1998.

[9] Kadambe, S. and Boudreaux-Bartels, G.F., 1992. Application of the wavelet transform for pitch detection of speech signals. IEEE transactions on Information Theory, 38(2), pp.917-924.

[10] Rabiner, L., 1977. On the use of autocorrelation analysis for pitch detection. IEEE transactions on acoustics, speech, and signal processing

[11] Madzarov, G., Gjorgjevikj, D. and Chorbev, I., 2009. A multi-class SVM classifier utilizing binary decision tree. Informatica, 33(2).

[12] B. Schuller, G. Rigoll, M. Lang, Speech emotion recognition combining acousticfeatures and linguistic information in a hybrid support vector machine-belief networkarchitecture, in: Proceedings of the ICASSP 2004, vol. 1, 2004, pp. 577–580.