Assignment Presentation

Applied Machine Learning

Student Name

FT-2017 Batch, Reg. No.: 17ETCS075004

M. Tech. in Machine Learning and Intelligent Systems

Module Leader:	**Dr. Subarna Chatterjee**
Module Name:	**Applied Machine Learning**
Module Code :	**MIS504**

Marking Scheme

Head	Maximum	Score
Technical Content	5	
Grasp and Explanation	5	
Quality of Slides and Delivery	5	
Q & A	5	
Total	**20**	

Presentation Outline

- Speech processing methods are preferred over Image processing techniques for Emotion Recognition.
- Algorithm for audio surveillance in PYTHON.
- Algorithm to recognize human emotion using speech processing.

Speech processing methods are preferred over Image processing techniques for Emotion Recognition.

- Emotions are fundamental for humans, impacting perception and everyday activities such as communication, learning and decision-making. They are expressed through speech, facial expressions, gestures and other non-verbal clues.
- **Parameters and challenges in speech detection for emotion analysis.**
  - Firstly, discovering which features are indicative of emotion classes is a difficult task. The key challenge, in emotion detection and in pattern recognition in general, is to maximise the between-class variability whilst minimising the withinclass variability so that classes are well separated. However, features indicating different emotional states may be overlapping, and there may be multiple ways of expressing the same emotional state. One strategy is to compute as many features as possible. Optimisation algorithms can then be applied to select the features contributing most to the discrimination while ignoring others, creating a compact emotion code that can be used for classification. This avoids making difficult a priori assumptions about which features may be relevant.
  - Secondly, previous studies indicate that several emotions can occur simultaneously. For example, co-occurring emotions could include being happy at the same time as being tired, or feeling touched, surprised and excited when hearing good news. This requires a classifier that can infer multiple temporally co-occurring emotions.
  - Thirdly, real-time classification will require choosing and implementing efficient algorithms and data structures.
  - Fourth, in persuasive communication, special attention is required to what non-verbal clues the speaker conveys. Untrained speakers often come across as bland, lifeless and colourless. Precisely measuring and analysing the voice is a difficult task.

Critical analysis of techniques in speech recognition for emotion analysis

Ref	Types of classifier	Types of features	Recognition Rate	Type of Dataset	Methods
H. Cao(2015)	SVM	Prosodic and spectral features	44.40%	Berlin & LDC & FAU Aibo dataset	Ranking SVM
T. L. New (2003)	HMM	Log frequency power coefficients (LFPC), MFCC	Average and best result 78% and 96% respectively	Two private speech dataset	Discrete HMM and LFPC to characterize speech signal
J.-H. Yeh(2011)	k-NN	Jitter, shimmer, formants, LPC, LPCC, MFCC, LFPC, PLP, and Rasta-PLP, SFS and SBS	Best 86%	Chinese emotional speech corpus We invited 18 males and 16 females	Segment based method by employing k-NN, SFS (sequential forward selection), SBS (sequential backward selection)
C.-C. Lee (2009)	Bayesian Logistic Regression, SVM	large-margin feature	70.1% & 65.1% for two and five class	AIBO dataset	Hierarchical structure for binary decision tree

Stance taken with justification

- Various speeches emotional recognition systems reviewed and discussed based on different approaches. Here the performance is also compared in terms of classifier, features, recognition rate, and datasets. Well-design classifiers have obtain high classification accuracies between different types of emotions. In this study HMM with adopting short time LFPC as a feature proves a good accuracy on different levels in the chart. Basharirad(2017) The majority of the current datasets are not capable for evaluation of speech emotion recognition. In most of them, it is hard even for human to specify different emotion of certain collected utterances; e.g. the human recognition accuracy was 80% for Berlin.
- **Conclusion**
  - From the above statements it is clearly evident that the facial expression does not exhibit as many classes of the emotions that is exhibited in voice/speech of a person.

Develop and simulate an algorithm for audio surveillance in PYTHON.

- Comparison of Audio Surveillance Algorithms.

Paper	Atypical sound classes	Model adaptation	Classifier	Features	Database
Ntalampiras et al. (2009)	Scream, gunshot, and explosion	MAP adaptation of GMMs	GMM	MFCC, MPEG-7, CB-TEO, Intonation	Large audio corpora from professional sound effects
Valenzise et al. (2007)	Scream and gunshot	-	GMM	Temporal, spectral, cepstral, correlation	Movie soundtracks, Internet, and people shouts
Radhakrishnan & Divakaran (2005)	Banging and nonneutral speech	-	GMM	MFCC	Elevator recordings
Clavel et al. (2005)	Gunshot	-	GMM	MFCC, spectral moments	CDs for the national French public radio
Rouas et al. (2006)	Shout	Adaptive threshold for sound activity detection	GMM, SVM	Energy, MFCC	Recorded during four scenarios
Vacher et al. (2004)	Scream and glass break	-	GMM	Wavelet based cepstral coefficients	Laboratory recordings and RCWP
Atrey et al. (2006)	Shout	-	GMM	ZCR, LPC, LPCC, LFCC	Recorded in office corridor
Ito et al. (2009)	Glass clash, scream, fire cracker	Adaptive threshold for abnormal sound event detection	GMM	MFCC, Power	Recorded under laboratory conditions

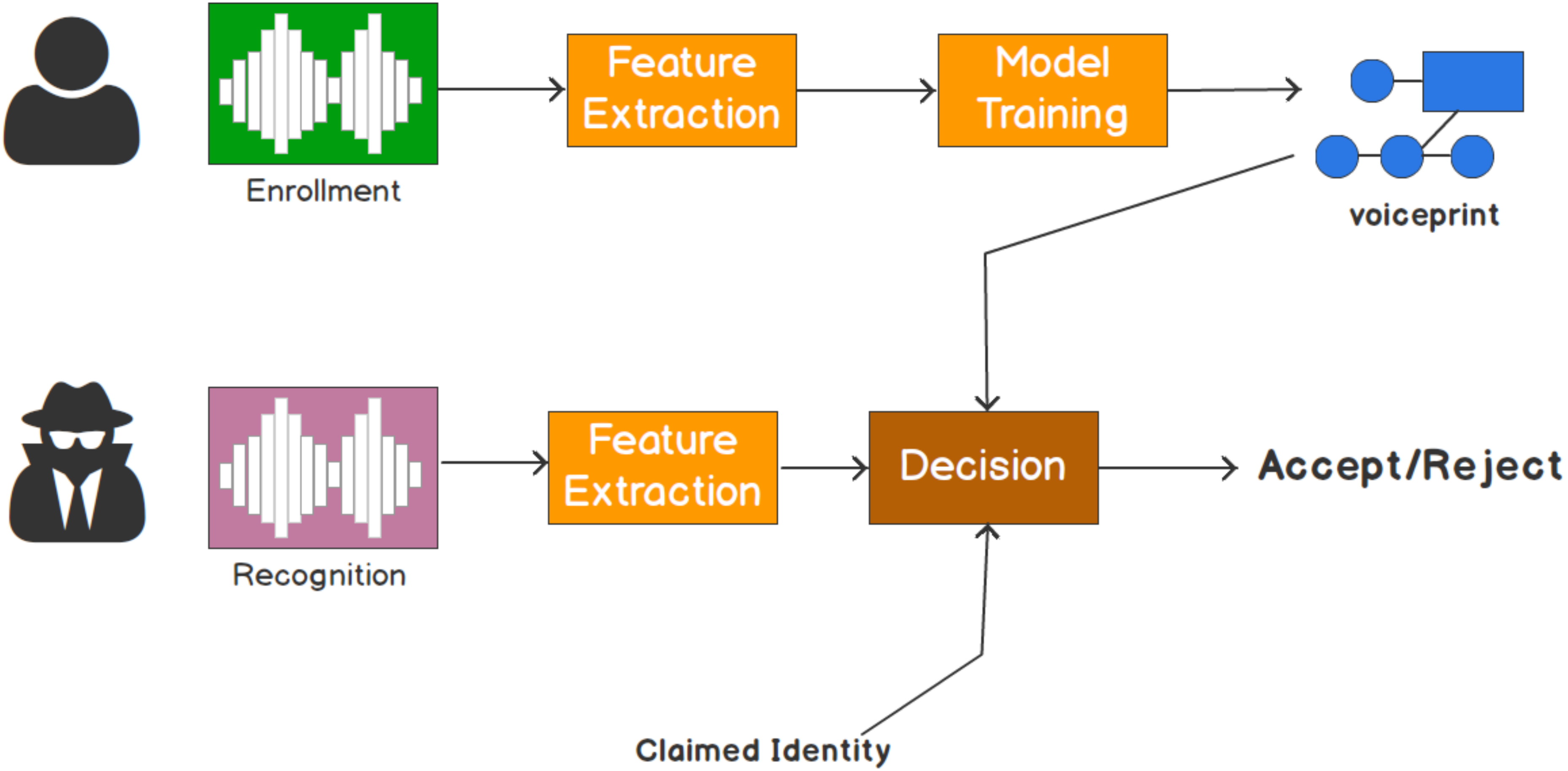
feature extraction techniques for Speech Recognition

Technique	Characteristics	Advantages	Disadvantages
LINEAR PREDICTIVE CODING	*Provides autoregression based speech features. *Is a formant estimation technique . *A static technique. *The residual sound is very close to the vocal tract input signal.[7]	*Is a reliable, accurate and robust technique for providing parameters which describe the timevarying linear system which represent the vocal tract. * Computation speed of LPC is good and provides with accurate parameters of speech. * Useful for encoding speech at low bit rate.	Is not able to distinguish the words with similar vowel sounds . * Cannot represent speech because of the assumption that signals are stationary and hence is not able to analyze the local events accurately. *LPC generates residual error as output that means some amount of important speech gets left in the residue resulting in poor speech quality.
MEL – FREQUENCY CEPSTRUM (MFCC)	* Used for speech processing tasks. *Mimics the human auditory system * Mel frequency scale: linear frequency spacing below 1000Hz & a log spacing above 1000Hz.	* The recognition accuracy is high. That means the performance rate of MFCC is high. * MFCC captures main characteristics of phones in speech. * Low Complexity.	*In background noise MFCC does not give accurate results. * The filter bandwidth is not an independent design parameter * Performance might be affected by the number of filters.
RelAtive SpecTrAl (RASTA Filtering)	* Is a band pass filtering technique. * Designed to lessen impact of noise as well as enhance speech. That is, it is a technique which is widely used for the speech signals that have background noise or simply noisy speech.	*Removes the slow varying environmental variations as well as the fast variations in artefacts. * This technique does not depend on the choice of microphone or the position of the microphone to the mouth, hence it is robust. * Captures frequencies with low modulations that correspond to speech.	*This technique causes a minor deprivation in performance for the clean information but it also slashes the error in half for the filtered case. RASTA combined with PLP gives a better performance ratio.
Probabilistic Linear Discriminate Analysis (PLDA)	*Based on i-vector extraction. The ivector is one which is full of information and is a low dimensional vector having fixed length. * This technique uses the state dependent variables of HMM. * PLDA is formulated by a generative model.	* Is a flexible acoustic model which makes use of variable number of interrelated input frames without any need of covariance modelling. * High recognition accuracy	* The Gaussian assumption which are on the class conditional distributions. This is just an assumption and is not true actually. * The generative model is also a disadvantage. The objective was to fit the date which takes class discrimination into account.

Classification techniques for Speech Recognition.

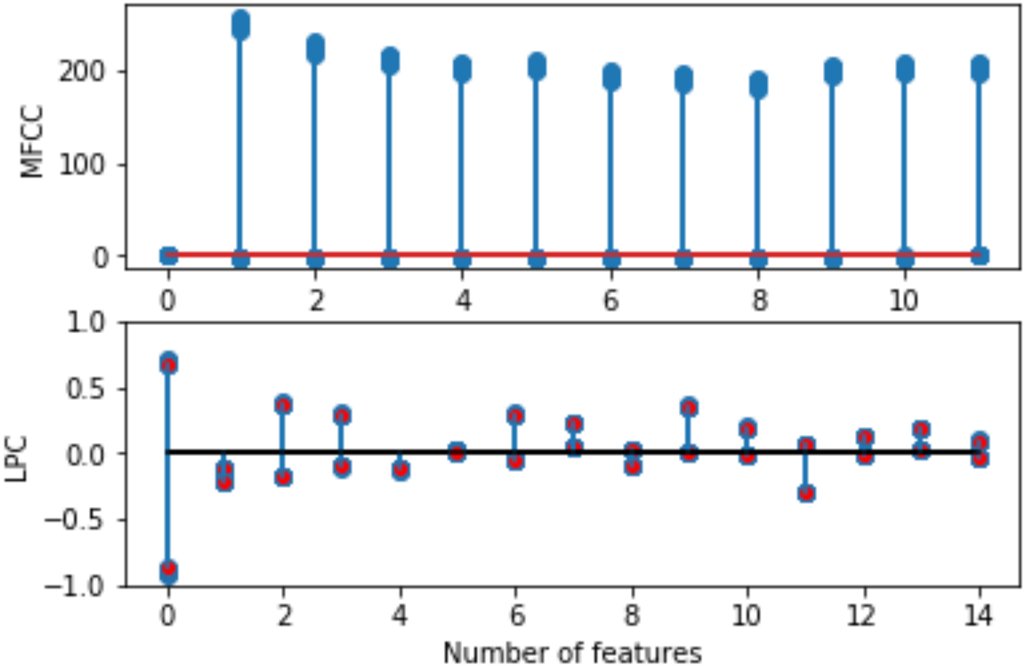
Technique	Characteristics	Advantages	Disadvantages
Guassian Mixture Model	Unsupervised	Needs less trainning and test data	Compromise between DTW and HMM
Dynamic Time Warping (DTW)	Unsupervised	Requires less storage space,beneficial for variable length	Cross-channel issue
Hidden Markov Model (HMM)	Unsupervised	Rail system outputs,efficient performance	Computationally more complex,more storage space
Vector Quantization(VQ)	Unsupervised	Computationally less complex	Real time encoding is complex
Support Vector Machine (SVM)	Supervised	Simple operation	binary SVM has limitations in speaker recognition

Audio surveillance system



Input and Their Features

File name	Voice	Uttered word
S1.wav	Female (Computer Generated)	Zero
S2.wav	Female (Computer Generated)	Zero
S3.wav	Female (Computer Generated)	Zero
S4.wav	Female (Computer Generated)	Zero
S5.wav	Female (Computer Generated)	Zero
S6.wav	Female (Computer Generated)	Zero
S7.wav	Female (Computer Generated)	Zero
S8.wav	Female (Computer Generated)	Zero



Validation of the speaker recognition system

True Speaker	Identified as (MFCC)	Identified as(LPC)
S1	S1	S5
S2	S8	S2
S3	S5	S1
S4	S6	S4
S5	S5	S5
S6	S3	S1
S7	S8	S8
S8	S8	S8
	Accuracy=37.5%	Accuracy = 50%

Develop an algorithm to recognize human emotion using speech processing

- **Choice of an appropriate emotion detection algorithm and its justification for the speech files given**
- This system consists of 5 steps, namely
  - 1.Emotional speech input
  - 2.Feature extraction and selection,
  - 3.Training,
  - 4.Classification,
  - 5.Emotion recognition
- **Justification for feature extraction:** MFCC (Mel frequency cepstral coefficient): Mel frequency cepstral coefficient is parametric representation known as voice quality feature, widely used in the area of speech emotion recognition. It provides better rate of recognition in both speech and emotion recognitions.
- **Justification for Classification:** The highly optimised SVMs produce higher cross validation accuracies than other algorithms. SVM classifier is binary decision algorithm and classification is mainly dependent on the MFFC feature.

Testing and validation of the reference model with the given speech file

- The provided Data set is been tested over the model and the result of the software programme are as follows.
  - Input File: ./new_test_sounds/YAF_young_angry.wav | Predicted Emotion Is: angry
  - Input File: ./new_test_sounds/YAF_young_disgust.wav | Predicted Emotion Is: disgust
  - Input File: ./new_test_sounds/YAF_young_fear.wav | Predicted Emotion Is: fear
  - Input File: ./new_test_sounds/YAF_young_happy.wav | Predicted Emotion Is: happy
  - Input File: ./new_test_sounds/YAF_young_neutral.wav | Predicted Emotion Is: neutral
  - Input File: ./new_test_sounds/YAF_young_ps(Pleasant Surprise).wav | Predicted Emotion Is: surprise
  - Input File: ./new_test_sounds/YAF_young_sad.wav | Predicted Emotion Is: disgust

Audio file	Emotion	Predicted Emotion
YAF_young_angry.wav	Angry	Angry
YAF_young_disgust.wav	Disgust	Disgust
YAF_young_fear.wav	Fear	Fear
YAF_young_happy.wav	Happy	Happy
YAF_young_neutral.wav	Neutral	Neutral
YAF_young_ps(Pleasant Surprise).wav	Surprise	Surprise
YAF_young_sad.wav	Sad	disgust
		Accuracy=85.71%

References

- Avetisyan, H., Bruna, O. and Holub, J. (2016). Overview of existing algorithms for emotion classification. Uncertainties in evaluations of accuracies. *Journal of Physics: Conference Series*, 772, p.012039.
- Ntalampiras, S. (2015). Audio Pattern Recognition of Baby Crying Sound Events. *Journal of the Audio Engineering Society*, 63(5), pp.358-369.
- Ntalampiras, S., Potamitis, I. and Fakotakis, N. (2009). An Adaptive Framework for Acoustic Monitoring of Potential Hazards. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009, pp.1-15.
- Pfister, T. (2010). *Emotion Detection from Speech*. Ph.D. Gonville & Caius College.
- Xia, S., Wang, J., Wang, R. and Zhao, L. (2015). An Improved Algorithm of Speech Emotion Recognition. *International Journal of u- and e- Service, Science and Technology*, 8(12), pp.217-226.