Shri Vile Parle Kelvani Mandal's

DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING

Approved by AICTE and Affiliated to the University of Mumbai

Department of Electronics & Telecommunication Engineering

Mini Project Report On

Title: **Gender Classification based on Pitch and MFCC using Support Vector Machine**

SUBMITTD BY:

Sr No.	Name of Students	SAP ID
1.	Viren Baria	60002160005
2.	Bhargav Desai	60002160017
3.	Sanjeet Krishna	60002160050
4.	Atulya Kumar	60002160052
5.	Parth Mehta	60002160063

Teacher's Name: Sunil Karamchandani

Shri Vile Parle Kelvani Mandal's

DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING

Approved by AICTE and Affiliated to the University of Mumbai

Department of Electronics & Telecommunication Engineering

CERTIFICATE

This is to certify that M/S. _____,

SAP ID_____ of TE EXTC 1: has submitted their

Case Study for Subject Name for the Academic Year 2018-2019.

Guide Examiner

Head of Department

EXTC Department

Index

Introduction:

A lot of approaches for classifying gender, require there be many auxiliary features like gait, behaviour, gesture, body shape along with the core biological features that distinguishes a male from a female like facial information and voice. This is understandable as facial information may not always be available and a classifier based on voice alone might work for one distribution and not for the other.

Determining a person's gender as male or female, based upon a sample of their voice seems to initially be an easy task. Often, the human ear can easily detect the difference between a male or female voice within the first few spoken words. However, designing a computer program to do this turns out to be a bit trickier.

Data Pre-processing:

In order to analyse gender by voice and speech, a training database was required. A database was built using thousands of samples of male and female voices, each labelled by their gender of male or female (male =1 and female=0). Voice samples were collected from the following resource:

http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Audio/Main/16kHz_16bit/

Around 3200 audio samples were downloaded by using python libraries like "requests" and "beautiful soup" to scrape audio data from the source. Once downloaded, metadata provided by the website was used to classify the audio samples into Male and Female.

```
User Name:23yipikayeSpeaker Characteristics:Gender: MaleAge Range: AdultLanguage: EN
Pronunciation dialect: OtherRecording Information:Microphone make: n/aMicrophone type:
Studio micAudio card make: unknownAudio card type: unknownAudio Recording Software:
VoxForge Speech Submission ApplicationO/S:File Info:File type: wavSampling Rate: 48000
Sample rate format: 16Number of channels: 1
```

Feature Extraction:

- Pitch

Pitch is a perceptual property of sounds that allows their ordering on a frequency-related scale or more commonly, pitch is the quality that makes it possible to judge sounds as "higher" and "lower" in the sense associated with musical melodies. Pitch can be determined only in sounds that have a frequency that is clear and stable enough to distinguish from noise. Pitch is a major auditory attribute of musical tones, along with duration, loudness, and timbre. The sensation of a frequency is commonly referred to as the pitch of a sound. A high pitch sound corresponds to a high frequency sound wave and a low pitch sound corresponds to a low frequency sound wave.

The pitch of a given audio file by first autocorrelating with a delay and then finding the position of the first peak.

- MFCC

In sound processing, 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.

MFCCs is derived as follows:

1. Take the Fourier transform of (a windowed excerpt of) a signal (setting n_mfcc = 10, hop_length=4000)
2. Map the powers of the spectrum obtained above onto the mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the mel frequencies.
4. Take the discrete cosine transform of the list of mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.

	pitch	mfcc1	mfcc2		mfcc108	mfcc109	mfcc110	label
0	111.8881	-510.598	-505.053		-5.70688	-6.48791	2.364354	male
1	53.51171	-505.605	-486.197		20.51213	12.87324	16.50841	male
2	125	-437.256	-441.068	----------	2.408163	6.308542	-12.4279	male

The features for all 3200 audio samples are thus extracted and we obtain a 3200 x 111 matrix saved as a CSV file where each sample consist of 110 mfcc values and 1 pitch value.

Machine Learning:

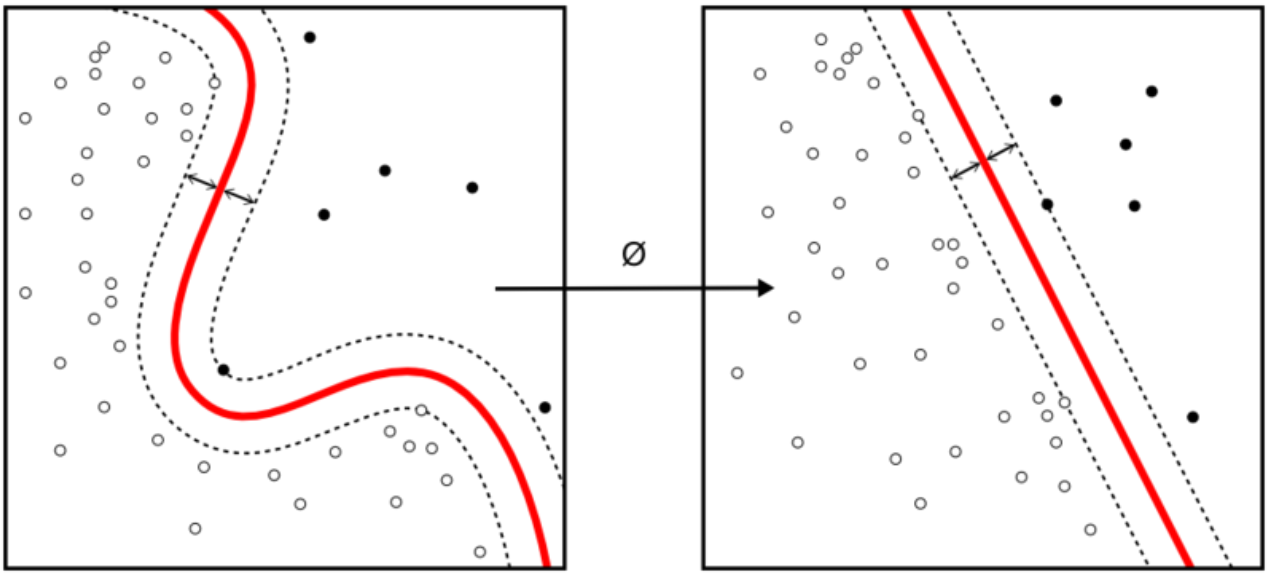
The classification machine learning algorithm used is Support Vector Machines (SVM) with radial basis function as kernel.

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right)$$ or

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma\|\mathbf{x} - \mathbf{x}'\|^2) \quad where\ \gamma = \frac{1}{2\sigma^2}$$

A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

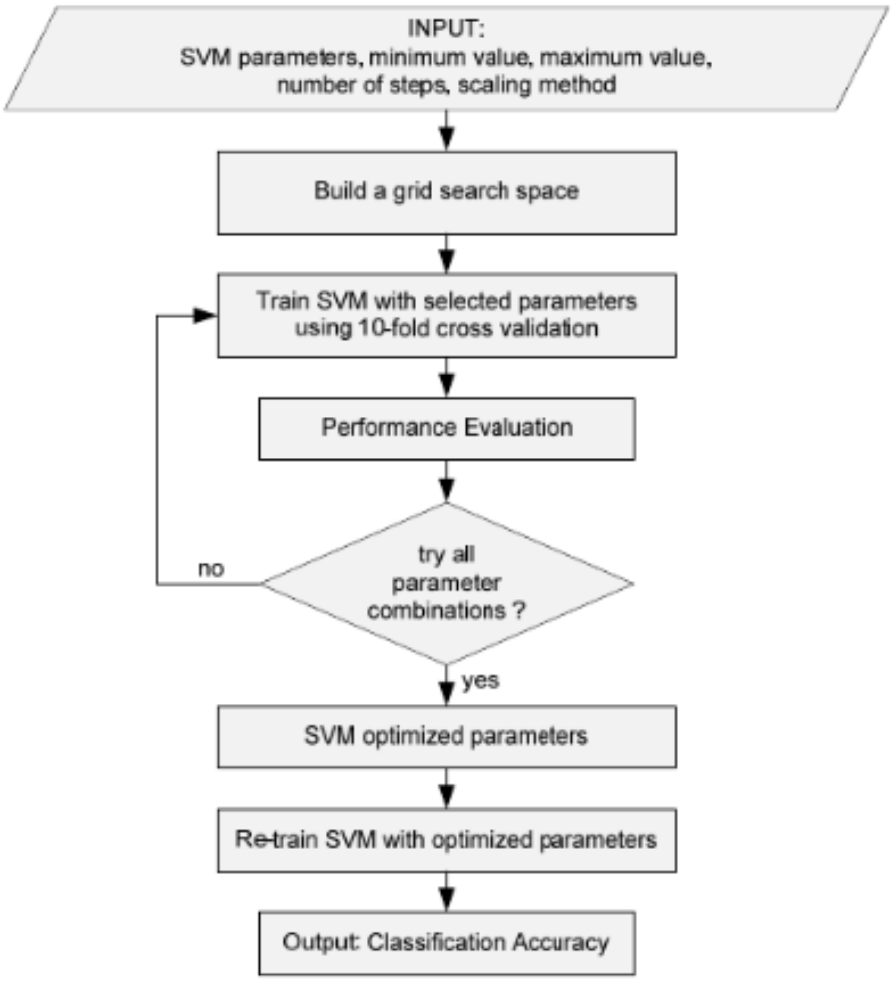
In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.



More formally, a support-vector machine constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outliers detection. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (functional margin), since the larger the margin, the lower the generalization error of the classifier.

Optimisation:

Without hyperparameter tuning. Classification accuracy turned out to be 74%. To obtain a better model accuracy grid search optimisation was required to tune the hyperparameters gamma and c.



The final values of tuned hyperparameters are C=3 and gamma=0.2 after grid search optimisation.

Conclusion:

We have successfully implemented and tested a robust machine learning model that can be used to classify any human voice irrespective of intrinsic features like accent, language, manner of speech and noise.

The accuracy for classification is 97.6% on test data.

References:

1. Trevor Hastie, Robert Tibshirani, Jerome Friedman. "The elements of Statistical Learning", p. 134.

2. Press, William H.; Teukolsky, Saul A.; Vetterling, William T.; Flannery, Brian P. (2007). "Section 16.5. Support Vector Machines". Numerical Recipes: The Art of Scientific Computing (3rd ed.). New York: Cambridge University

3. Digital Processing of Speech Signals, Rabiner,LR

4. Kumar, R. (2019). Machine Learning Quick Reference. Birmingham: Packt Publishing Ltd.