AGE ESTIMATION FROM SPEECH USING GAUSSIAN MIXTURE MODEL

Jyoti Gupta
GLBITM,
Greater Noida
Shagufta
GLBITM,
Greater Noida
Neha
GLBITM,
Greater Noida
Poornima Gupta
GLBITM,
Greater Noida

ABSTRACT - This project, Age Estimation from speech is done to find out the age of a person by speech. There are many features present in the speech of a person, from that we are using spectral features and based on that we are doing our classification. There are several classifiers such as Artificial Neural Network, Hidden Markov Models, Gaussian Mixture Models and from all we are using Gaussian mixture model. Gaussian mixture models have been found to perform good with MFCC and therefore we are using this as a classifiers.

General Terms
Pattern recognition: Pattern recognition is a branch of machine learning that focuses on the recognition of patterns and regularities in data.

Mel frequency cepstral coefficients: The Mel-frequency Cepstrum 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.

Gaussian Mixture Model: A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. One can think of mixture models as generalizing k-means clustering to incorporate information about the covariance structure of the data as well as the centers of the latent Gaussian.

Keywords - Age approximation, Mel frequency cepstral coefficients, Spectral features, Age approximation models, Gaussian mixture model.

I. INTRODUCTION

One of the most easiest, common and important form of communication we use is speech. Speech has many information like about the language of the speaker, gender of the speaker, age of the speaker and many more. All the information play important role in many applications and the information about the age of the speaker is our concern. It can be used for verification of a person in sports or in authentication to adult websites, in voting also it can be used for verification of the voter.

It is very difficult to find the correct age of a person so we focus on the approximation of the age of a person. There are many features that are present in the speech and from that features we are extracting that feature which is best suited for the identification of the age. One such feature that is found to be most successful is spectral feature. Spectral features mainly have vocal tract information like format frequencies, sequential variation in the size and shape of vocal tracts, spectral roll off, spectral bandwidth and etc.

Many spectral features has been used for identifying emotion. Now there are many feature like MFCC, LPCC, BFCC but among all MFCC has been used. Usually feature extraction is done through block processing approach. Full speech signal is handled frame by frame. These frames are of size 20ms approx. In this speech signal is assumed to be stationary. Finally, Mel frequency cepstral coefficient (MFCC) are extracted as spectral feature and further process of age estimation is done. MFCC has six main stages that will be discussed later.

We need to collect the data and store it in database and on that we perform the classification of age according to the defined age group, which sample will go in which category.

II. FEATURE EXTRACTION

Processing of speech is one of the important application area of DSP(digital signal processing). Fields involve in speech processing are speech recognition, speaker recognition, speech synthesis, speech coding etc. The main objective of automatic speaker recognition is to extract, characterize and recognize the information related to identity of speaker. Speech processing has an important application in telephone communication, domestic appliances control, voice dialing, call routing, speech to text conversion, text to speech conversion etc. In modern era speech processing has been develop as a novel approach in security. Feature extraction is the first step of speech recognition. Algorithm used for this are MFCC (Mel Frequency Cepstrum Coefficient), LPC (Linear Predictive Code), PLP (Perceptual Linear Prediction). MFCC and PLP is based on nature of
speech while it extract the feature, however LPC predict
the future feature based on previous feature.

The human speech contain various features that can be
used to identify speakers. Speech contains significant
energy from zero frequency up to around 5 kHz. Speech
signal property changes as a function of time which is
remarkable. Time varying Fourier transformation is
used to study the spectral properties of speech signal.
However temporal properties (correlation, energy etc)
assumed to be constant over a short period. Speech
signal is divided into number of blocks of short duration
using hamming window so that normal Fourier
transform can be used.

The most dominant method used to extract spectral
features is calculating MFCC. MFCC technique is used
in speech recognition based on frequency domain Mel
scale. Mel scale is based on human ear scale. Frequency
domain features are more accurate than time domain
features. The extracted MFCC’s features are quantized
to a number of centroid using vector quantization
algorithm. These centroids constitutes the code book of
speaker. Feature of MFCC are calculated in training and
testing phase. Speakers uttered the same word in both
the phase. The euclidean distance between the MFCC’s
of each speaker in training phase to the centroids of
individual speakers in testing phase is measured and
according toh the minimum Euclidean distance the
speaker is identified. The code is developed in the MAT
LAB environment.

MFCC is a representation of a real cepstral of a short
tome signal obtained from fast Fourier transform of the
signal. The main difference from real cepstral is that in
this non linear frequency scale is used which
approximates the auditory system behavior. These
coefficient are robust reliable to variations according to
speakers. MFCC is an audio feature extraction
technique in which parameters are extracted from
speech similar to ones that are used by humans for
hearing speech while reemphasizing all other
information at the same time.

III. SPEAKER RECOGNITION

Anatomy of vocal tract is used in recognition of
speaker. Anatomical structure of the vocal tract is
unique for every person. Hence the voice signal of each
individual differ which is helpful in identifying the
speaker. Anatomical structures are intrinsic property,
voice comes under bio metric identity. Speaker
recognition system involve training and testing.
Training is the process of familiarizing the system with
voice characteristics of the speaker. Testing is the actual
recognizing task. Speech signal can be represented by a
sequence of feature vector. Feature selection and
extraction is the selection of appropriate features along
with the method to estimate them.

IV. SPEAKER RECOGNITION

Techniques

Speaker recognition concentrate on recognizing the
unknown speaker from a set of known speakers. Speaker
recognition system consist of four main part.

Processing of front end: In this samples speech signal
is converted into feature vector set which characterize
the speech properties that can separate different
speakers. Speaker modeling: In this part by modelling
the distribution of feature vector feature data is reduced.

Speaker database: Speaker models are stored in this.

Decision logic: Makes the final decision about speaker
identity by comparing unknown feature vectors to all
models. Following are the techniques of Feature
Extraction:

4.1 LPC (Linear Predictive Codes):

LPC is desirable to compress signal for efficient
transmission and storage. LPC analyzes the speech
signal by estimating the formats, removing their effect
(reverse filtering) from speech signal and then estimate
the intensity and frequency of remaining buzz (residue).
In LPC each sample signal is expressed as a linear
combination of previous sample.

4.2 PLP (Perceptual Linear Prediction):

In this human speech is modeled on the concept of
psycho physics of hearing. PLP improves speech
recognition rate by discarding irrelevant information of
the speech. It differ from LPC in the fact that its
spectral characteristics have been transformed to match
characteristics of human auditory system.

4.3 MFCC(Mel Frequency Cepstral Coefficient):

The important task in the design of any speech
recognition system is the extraction and selection of
best parametric representation of acoustic signal. It is
provided by MFCC. MFCC is the result of the cosine
transform of real logarithm of short term energy
spectrum expressed on a Mel frequency scale. The
calculation of MFCC includes:

Mel frequency wrapping: Pitch is measured on scale
called Mel scale. The Mel frequency scale is a linear
frequency spacing below 1000 Hz. Formula to compute
the Mel for a given frequency f in Hz:

\[ \text{Mel}(f)=2595 \log_{10}(1+f/700). \]

Cepstrum: This is the final in which log Mel spectrum is converted back to
the time. The result is called Mel frequency cepstral
coefficient (MFCC). The cepstral representation of
the speech spectrum provides a fine representation of the
local spectral properties of the signal for a given frame.
We can convert Mel spectrum coefficient to the time
domain using DCT (discrete cosine transform) and
finally log Mel spectrum is converted back to the time
domain.

LPC parameter is not so efficient because of its linear
computation nature. As human voice is nonlinear in
nature, LPC is not a good choice. PLP and MFCC are
derived on the concept of logarithmically filter bank
with concept of human auditory system and hence is
better as compared to LPC.

V. GAUSSIAN MIXTURE MODELS

Mixture models are a type of density model which
comprise a number of component functions, usually
Gaussian. These component functions are combined to
provide a multi modal density. There are different
classifiers such as Hidden Markov Models, Gaussian
Mixture Models, K-nearest neighbor algorithm,
Decision trees, Kozinees algorithm, Non-linear
(smooth) SVM, Polynomial classifiers, Perception
classifiers, Logistic Recognition, Least square methods
which can used for different speech processing tasks,
such as speaker recognition, speech recognition,
speaker verification, emotion classification, age
approximation and so on. Gaussian Mixture Models is
one of the most statistically matured methods for
clustering and density estimation. MFCCs is used as
classifiers, which is used with Gaussian Mixture Model.
GMMs helps in capturing the distribution of data points
from the input feature space. In GMM modelling
probability density function of input data point is done
using multivariate. Expectation maximization algorithm
is used for refining the weight which find the maximum
likelihood parameters of a statistical model. The
number of clusters into which data points in GMM are
classified is the number of components. In order to get
more generalized clusters, large data set should be taken
during training of data set in GMM.

Number of components is defined as the number of
Gauss in the mixture model. These component capture
finer level details among the feature vectors of each
emotion. Number of components indicate the number of
clusters.

VI. REFERENCES

[1] Shasidhar G. Koolagudi, Reddy, R. ,Yadav, J. and
corpus for emotion analysis, IEEE International
Conference on Devices and Communications.
Fundamentals of Speech Recognition. Englewood
Cliffs, New Jersey: Prentice-Hall.
of vocal disguise on estimations of speaker’s ages.
Perceptual and Motor Skills, 45 (3), 1311-1315.
for speaker recognition ,” MS thesis, Department of
Computer Science and Engineering, Indian Institute of
Technology Madras, Chennai600 036, India.
speaker Recognition,” MS thesis, Department of
Computer Science and Engineering, Indian Institute of
Technology Madras, Chennai 600 036, India.
Sreenivasa Rao,” Emotion Recognition Using LP
Residual, Proceedings of the 2010 IEEE Students
Technology Symposium, 3-4 April 2010,IT Kharagpur.
F. 2014, A Systematic Analysis of Automatic Speech
Recognition: An Overview. International Journal of
Current Engineering and Technology, Vol.4, No.3