Multi-Knowledge ELECTRONIC COMPREHENSIVE JOURNAL FOR EDUCATION AND SCIENCE PUBLICATIONS (MECSJ)

ISSUE (13) , Oct (2018)

ISSN : **2616-9185**

www.mecsj.com

Emotion Detection from Speech Emotion Detection from Speech

Mazin Riyadh AL-Hameed

Master of computer science, Modren University for business & science, Lebanon

Cutemazen2008@gmail.com

00961-70129239

Abstract

Speech recognition applications are becoming useful more and more nowadays. Various aware applications of interactive speech are available in the market. But usually they are executed and meant for the traditional computers general-purpose. With needs growth for embedded computing and the demand for platforms emerging embedded, the speech recognition systems is required (SRS) to be available too on them. PDAs and other handheld devices are becoming powerful and affordable more and more as well. Running multimedia is become possible on these devices. As efficient alternatives for such devices Speech recognition systems emerge where typing attributed to their small screen limitations becomes difficult.

The obvious voice recognition technology advantage is for people who have various circumstances disabilities, who have difficulty in several cases expressing themselves.

Multi-Knowledge ELECTRONIC COMPREHENSIVE JOURNAL FOR EDUCATION AND SCIENCE PUBLICATIONS (MECSJ)

ISSUE (13) , Oct (2018)

ISSN : **2616-9185**

www.mecsj.com

I- Introduction

Although, speech emotion detection is a new field of research relatively, it has many applications potential. In human-human or human-computer interaction systems, systems of emotion recognition could provide improved services to users by being to their emotions adaptive. Virtually, more interaction of realistic avatar could have help in simulation with emotion recognition.

The detecting body of emotion in speech work is quite limited. Currently, researchers are still debating the emotion in speech recognition are influenced with what features. There is also considerable uncertainty as to the best classifying emotion algorithm, and classing together which emotions.

In this project, using Support Vector Machines (SVMs) and K-Means to classify opposing emotions, these issues are attempt to be addressed in this project. We separate the speaker gender speech to investigate the relationship between emotional content of speech and gender.

From human speech, we can extract a variety of spectral features and temporal. Using statistics relating to Mel Frequency Cepstral Coefficients (MFCCs), Formants of speech and the pitch as inputs to algorithms classification. The recognition of emotion accuracy of these experiments allows explaining to us, the most emotional information is carried with which features and why.

Classing emotions together is also handled by allowing us to develop criteria for this. Also, achieving high recognition of emotion accuracy can be done using these techniques.

A-Motivation

Emotions are fundamental for humans, such as communication, the impact perception and everyday activities, decision-making and learning. Through speech, they are expressed, gestures, facial expressions and other non-verbal clues.

Multi-Knowledge ELECTRONIC COMPREHENSIVE JOURNAL FOR EDUCATION AND SCIENCE PUBLICATIONS (MECSJ)

ISSUE (13) , Oct (2018)

ISSN : **2616-9185**

www.mecsj.com

Detection of Speech emotion refers to vocal behavior analyzing as an affect marker, focusing, on speech nonverbal aspects. Basically, assumption is that there is a set of parameters objectively measurable in voice, reflecting the currently expressing affective state by a person. Supported by most affective states involve physiological reactions as a fact, this assumption modifies the process that produce voices. By example, anger increases tension in muscle and often produces changes in respiration, influencing the vibration of the vocal tract and vocal folds' shape and affecting the speech acoustic characteristics. So far, less attention is received by vocal emotion expression than the facial equivalent, mirroring the emphasis relative by pioneers such as Charles Darwin.

Previously, emotions were considered unmeasurable and computer scientists didn't study it consequently. Although recently, a contributions increase is received in this field, a new study area is remaining with a potential applications numbers. Including hearing aids emotional for autism people; angry caller detection at an automated call center to human transfer or style adjustment presentation of an e-learning computerized tutor when student is bored.

A new emotion detection application proposed in this dissertation is speech tutoring. Especially, in communication persuasive, special attention to what the speaker conveys non-verbal clues is required. Untrained speakers come across as bland often, colorless and lifeless. Precisely analyzing and measuring the voice is a very complicated task has been entirely subjective in the past. Using a similar approach as for emotions detecting, this report shows that such judgements can be made objective.

B-Challenges

Some of implementing expected challenges is described in this section, in a real-time detector of speech emotion.

Firstly, the difficult task is discovering which features are emotion classes indicative. Generally, in pattern recognition and in emotion detection, the challenge key is to maximize the variability between-class whilst minimizing the variability within class so that classes are separated well. However, different emotional indicating features states may be overlapping, and multiple ways of expressing may have the same emotional state. Compute as many features as possible is the only strategy. Algorithms optimization can be applied to the features selected, while ignoring others, to contribute most to the discrimination, creating an emotion compact code useful for classification. This avoids making a priori assumptions difficult about which features may be relevant.

Secondly, several emotions can occur simultaneously as indicated in previous studies. By example, emotions co-occurring can include being tired at the same time as being happy, or feeling surprised, excited and touched when hearing good news. A classifier is required here, that can infer multiple co-occurring temporally emotions.

Thirdly, classification in real-time will require implementing and choosing data structures and efficient algorithms.

Despite some of their working systems existing, implementations are generally expected to be imprecise and imperfect and are still seen as challenging.

II- **Speech recognition Overview**

The speech among human being is the primary mode of communication also it is the most efficient and natural exchanging information form among speech human. So, speech recognition for HCI (Human Computer Interaction) is logical to be natural language in the next technological development. Speech Recognition can be defined as converting speech signal process to a words sequence by means implemented Algorithm as a computer program. Processing of Speech is one of the signal processing exciting areas. The speech recognition goal area is developing system and technique for machine speech input. Based on advanced major in statically speech modeling, today, automatic speech recognition find application widespread in task requiring human machine interface such as automatic call processing.

Since the 1960s research of computer scientists is concentrated on the ways and means to make human recorded, interpreted and underattended by speech computers. Throughout, the decades this has been a daunting task. Even the problem most rudimentary like digitalizing (sampling) voice was a big challenge in early years. The first systems didn't arrive until the 1980s, which could decipher speech actually. Off course, these initial systems were in scope and power very limited. Among human being, the communication is dominated by spoken language, therefore it is natural to expect interfaces with computer speech for people.

As with all technology, the technology of speech recognition is most effective when combined in punctuation with direct skills instruction, paragraphing and sentence skills.

Multi-Knowledge ELECTRONIC COMPREHENSIVE JOURNAL FOR EDUCATION AND SCIENCE PUBLICATIONS (MECSJ)

ISSUE (13) , Oct (2018)

ISSN : **2616-9185**

www.mecsj.com

A- Speech

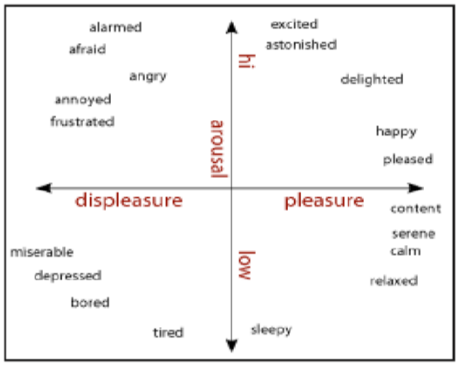
The primary communication means between humans is Speech. It is a complex signal that contain message information, language, emotional states, speaker and so on. In terms Speech is referred of the production and perception of sounds used in vocal language. Speech production defines the processing of the spoken words and the formulation of their phonetics whereas speech perception refers to process on which humans can understand and interpret the human language presented sound.

B- Emotions

Emotions are defined as psychological feeling and physical changes that influences thought of humans and behavior. It is associated with personality, temperament, motivation, mood, energy etc. The emotions can be categorized by dimensions grouping.

C- Dimensional Model of Emotions

Dimensional model conceptualizes emotions usually by defining them in 2 or 3 dimensions. All dimensional models are valence and arousal incorporate. In 1980, Circumflex of effect (Figure 1) is introduced by James Russell. He mapped in two dimensional axes emotions as high/low arousal in vertical axis and pleasure/displeasure in horizontal axis.



III- **Speech Recognition Techniques**

The speech recognition goal is the ability to the machine for hear, understand, and act upon information spoken. The earliest systems of speech recognition were attempted first in the early 1950s at Bell Laboratories, Balashek, Biddulph and Davis developed an isolated digit system Recognition for a speaker single. The automatic speaker recognition goal is to analyze and extract characterize and recognize the speaker identity information. The system of speaker recognition may be work in four stages:

D- Analysis:

The analysis of speech uses the vocal tract for signal extracting to recognize the speech. For segmenting the got signal there are suitable frames.

10-30 MS range: for vocal tract information extracting.

3-5 MS range: for extracted tract analyzing

Multi-Knowledge ELECTRONIC COMPREHENSIVE JOURNAL FOR EDUCATION AND SCIENCE PUBLICATIONS (MECSJ)

ISSUE (13) , Oct (2018)

ISSN : **2616-9185**

www.mecsj.com

E- Feature Extraction Technique:

The signal extracting technique in dimensionality reducing while maintaining the signal power.

Modeling Technique

Using specific features speaker models are generated by this technique. Classified into speaker identification and speaker recognition. The speaker recognition is separated into two Sections: speaker dependent and speaker independent.

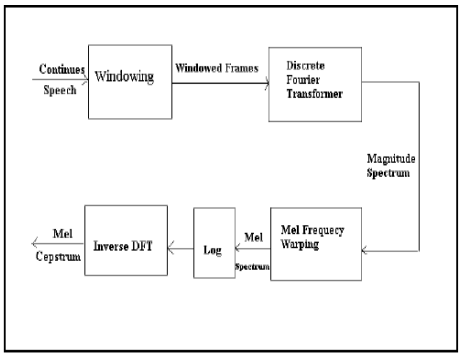
In the speech recognition for speaker independent mode of, the specific characteristics of speaker are ignored by the computer of the speech signal and extract the intended message.

The speech signal is compared by a speaker identification from an unknown speaker to a known speaker database. The speaker can be recognized by the system that has been trained with a number of speakers.

- Following are used speech recognition process modeling:

i. The acoustic-phonetic approach:

This approach was studied and used over 40 years' depth. It depends on the phonetics identification. By the similarity in phonetics some problems faced such as /d/ and /e/. It is not over computer applications widely known.



ii. Pattern Recognition approach:

Set of labeled training samples are established by this approach via a training formal algorithm in order to compare the sounds patterns, phrases or words. Between the unknown speeches, a direct comparison is made with each pattern learned possible in the stage of training. the best fits are written, by this comparison. In the last six decades, the approach of pattern-matching has become the method predominant for speech recognition.

iii. Template based approaches

The speech is compared to templates predefined to find the best matches. These templates are prototypical speech patterns collection, stored as patterns reference and representing the candidate's words dictionary. Recognition is then done by unknown utterance comparing to these patterns in order to find the pattern category

ISSN : **2616-9185**

that fits best. It has an advantage that due to segmentation errors are reduced. But on the other hand, each single word needs its specific template which is fixed, so any speech variation can only be modeled by the uses of more than one template for each word, and that is impractical.

iv. Dynamic Time Wrapping

Generally, DTW is a method allowing a computer for finding an optimal match between two given sequences (e.g. time series) with certain restrictions. Dynamic time warping (DTW) is such an approach typical for a speech recognition matching template based approach and also DTW compresses and stretches various utterance sections so as to find alignment resulting in best possible match between utterance and template on the basis of frame by frame. "Frame" means short segment (10-30ms)

v. Statistical based approaches

Although, studied and introduced initially in the early 1970s and late 1960s. In the last several years, statistical hidden methods of Markov models have become popular increasingly. This was because of two reasons:

- The models are very mathematical structure rich and hence can form the basis for use theoretical in a wide application range.
- The models work very well, when applied properly, in several applications practices.
- In General, Real-world processes produce out-puts observable, which can be characterized as signals.
- The signal in nature can be **discrete** – e.g. character from alphabet, from a codebook quantized vector [*].
- The signal in nature can be **continuous** – e.g. music, management of temperature, …
- The signal can be **stationary** – its properties statistical don't vary with time.
- The signal can be **nonstationary** - properties statistical can varies with time.
- The signal can be **pure** – strictly coming from a source single, or from another signal sources **corrupted** such as noise.

Multi-Knowledge ELECTRONIC COMPREHENSIVE JOURNAL FOR EDUCATION AND SCIENCE PUBLICATIONS (MECSJ)

ISSUE (13) , Oct (2018)

ISSN : **2616-9185**

www.mecsj.com

The basis for the description theoretical can be provided by a signal model of a signal processing unit which can be used for signal processing to provide an output desired.

Example, signal model is used to provide a noise removing system and from the signal source disturbances.

Another advantage, is that the signal source information can be known without having the source actually, costs too much when getting the source signal.

vi. Artificial Intelligence approach

The Artificial Intelligence approach is a mix of pattern recognition approach and the approach of acoustic phonetic. In this, it exploits the concepts and ideas of pattern recognition methods and Acoustic phonetic.

Speech recognition artificial intelligence (AI) involves two ideas basic. First, it involves studying the human beings thought processes. Second, it deals with processes via machines representing (like robots, computers, etc.).

vii. Stochastic Approach

Dealing with incomplete information is the main uses of this approach. Which comes from confusable speaker's variability, homophones and sounds.

Matching Techniques

i. Whole-word matching:

- The incoming digital-audio signal is compared by the engine against the word prerecorded template.
- Prerecord for every word is required by the user that will be recognized several hundred thousand words sometimes.
- Large amounts of storage also required by whole-word templates (between 50 and 512 bytes/word).
- Practical only on the knowing of the recognition vocabulary on developing the application.

ii. Sub-word matching:

- The engine looks for usually phonemes sub-words and then performs pattern recognition further on those.
- It takes more processing than matching the whole-word.
- This technique requires much less storage (between 5 and 20 bytes/word). In addition, the word pronunciation can be guessed, without requiring to speak the word beforehand from the user, from English text.

[*] **VQ** the Vector Quantization is nothing more than an approximator. It is similar to

Multi-Knowledge ELECTRONIC COMPREHENSIVE JOURNAL FOR EDUCATION AND SCIENCE PUBLICATIONS (MECSJ)

ISSUE (13) , Oct (2018)

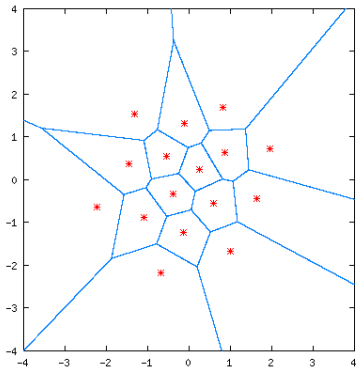
ISSN : **2616-9185**

www.mecsj.com

"rounding-off" (get to the nearest integer). An example of a VQ 1-dimensional is shown below in Figure 3:

Here, every less than -2 number is to -3 approximated. Every number between 0 and - 2 are to -1 approximated by. Every number between +2 and 0 are to +1 approximated. Every number greater than +2 is approximated to +3. Note that approximated values are represented uniquely by 2 bits. This is a VQ 1-dimensional, 2-bit. It has 2 bits per dimension rated.

Here, every falling pair of numbers in a region particular is approximated by a red associated star with that region. Knowing,



that there are 16 regions and 16 red stars, each one can be represented uniquely by 4 bits. Thus, this is a 4-bit VQ, 2-dimensional. It's also 2 bits per dimension rated.

In the two examples previous, code vectors were the name of red stars and the defined by the blue borders regions were called encoding regions. The all code vectors set is called the codebook and the all encoding regions set is called the space partition.

In this domain, the lack of accuracy in understanding words was one of the most difficulties faced by the system due to its towards noise sensitivity and the harmonies similarity of between words. Moreover, big role is played by the accent in the words identification.

IV- Corpus of Emotional Speech Data

The used data for this project comes from the Data Consortium's Linguistic study on Prosody Emotional and Transcripts Speech. The recordings audio and transcripts corresponding were over an eight-month period collected in 2000-2001 and are designed to support emotional prosody research.

The professional actor's recordings consist of reading a neutral utterance semantically series (numbers and dates) spanning 14 distinct emotional categories, selected after the study of Banse & Scherer's about expression of vocal emotional in German. There were five speakers' females and three speakers' males, all in mid-20s. The utterances number that belong to each category of emotion is shown in Table 1. With a sampling rate of 22050Hz the recordings were recorded and encoded in interleaved 16- bit PCM two-channel, high-byte-first ("big-endian") format. They were then to

Multi-Knowledge ELECTRONIC COMPREHENSIVE JOURNAL FOR EDUCATION AND SCIENCE PUBLICATIONS (MECSJ)

ISSUE (13) , Oct (2018)

ISSN : **2616-9185**

www.mecsj.com

single channel recordings converted by taking the both channels average and DC-offset removing .

Neutral	Disgust	Panic	Anxiety
82	171	141	170
Hot Anger	Cold Anger	Despair	Sadness
138	155	171	149
Elation	Happy	Interest	Boredom
159	177	176	154
Shame	Pride	Contempt	
148	150	181	

Table 1 - Number of utterances belonging to each Emotion Category

V- **Feature Extraction**

A-Pitch and related features

Pitch is a speech signal fundamental frequency F_0. In general, it represents the production of vocal cords vibration during the production of sound. The pitch signals usually have two pitch frequency characteristics and air glottal velocity. Given by the harmonics number the pitch frequency is presented in spectrum. The pitch estimation is one of the tasks complex. We use the estimation algorithm of YIN Pitch to detect values of pitch.

Bäzinger et al. argued that pitch related statistics conveys emotional status considerable information. Yu et al. have shown that some pitch statistics carries information about Mandarin speech emotion. For this

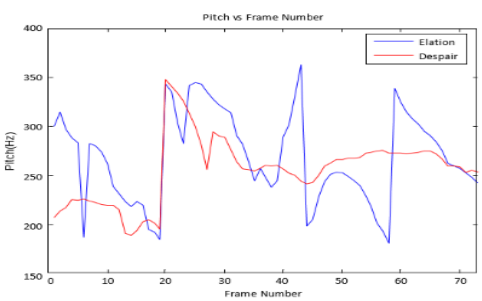


Figure 1 - Variation in Pitch for 2 emotional states

project, pitch is extracted from the waveform speech using a RAPT algorithm modified version to implement in the VOICEBOX toolbox the pitch tracking. Using a 50ms length frame, each frame pitch was calculated and placed in a vector that corresponds the frame. If it's an unvoiced speech the corresponding marker was set to zero in the pitch vector.

The variation for a female speaker in pitch is showed Figure 5 uttering "Seventy-one" in despair and elation emotional states. From this figure, it is evident that the variance and mean of the pitch is higher when uttering "Seventy-one" in elation rather than despair. For capturing these and other characteristics, it needs to calculate

Multi-Knowledge ELECTRONIC COMPREHENSIVE JOURNAL FOR EDUCATION AND SCIENCE PUBLICATIONS (MECSJ)

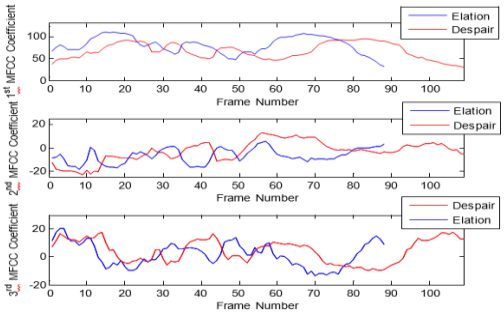
ISSUE (13) , Oct (2018)

ISSN : **2616-9185**

www.mecsj.com

the following statistics from the pitch and uses it in the pitch feature vector:

- Median, Mean, Variance, Minimum, Maximum (for the pitch vector and its derivative).
- Voiced average energies and unvoiced speech.
- Rate of speaking (average length



inverse of the utterance voiced part).

- Hence, the vector of pitch feature is 13-dimensional.

B-MFCC and related features

In speech recognition, Mel Frequency Cepstral Coefficients are the most widely used features. David and Murmelstein's who introduces these features, by MFCC processor main purpose is to mimic the human ears behavior. The MFCC main steps are identified (Khalifa et al., 2004) and the MFCC block diagram is shown in Figure 6.

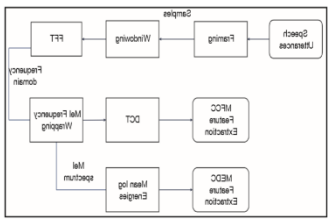


Figure 2 - MFCC and MEDC Feature Extraction

MFCC are also in many applications the most widely used speech spectral representation, including speaker recognition and speech. Kim et al. argued that MFCCs relating statistics carries also emotional information. For each speech frame of25ms length, thirteen standard parameters of MFCC are calculated by below steps:

- Taking the STFT absolute value.
- Warping the absolute value to a Mel frequency scale
- Taking the log-Mel- spectrum DCT and returning the first 13 components.

Figure 3 - Variation in MFCCs for 2 emotional states

The three MFCCs variation are shown in Figure 7 for a female speaker uttering in despair and elation emotional states "Seventy-one". It is evident from this figure that when uttered in elation rather than despair "Seventy-one" the mean of the first coefficient is higher, but for the second and third coefficients is lower. For capturing these and other characteristics, it needs to extract MFCCs based statistics. For each coefficient and its derivative, we calculated across all frames the variance, mean, minimum and maximum. We also

Multi-Knowledge ELECTRONIC COMPREHENSIVE JOURNAL FOR EDUCATION AND SCIENCE PUBLICATIONS (MECSJ)

ISSUE (13) , Oct (2018)

ISSN : **2616-9185**

www.mecsj.com

calculate of each coefficient mean and its derivative the variance, mean minimum and maximum. Each vector of MFCC feature is 112-dimensional.

Usually the input recording of speech is done through microphone with a sample rate of 16000 Hz. The MFCC calculating steps are described below:

iii. Framing

In this step, the continuous speech input is segmented into N frames sample. N samples composing the first frame, after N M samples composing the second frame, third frame contains 2M and so on. With 2040ms time length we frame the signal here. Therefore, the 16Khz signal frame length will have 0.025*16000 = 400 samples.

iv. Windowing

In this step, the process is used in order to remove the start and end discontinuities by windowing each individual frame. Remove distortion is mostly done by hamming window.

v. Fast Fourier Transform

FFT algorithm is used here to convert from time domain the N samples to frequency domain. It is used for speech frequency spectrum evaluation.

vi. Mel Filter Bank

Each frequency is mapped from frequency spectrum in this step maps to Mel scale. Which usually consist of triangular filters overlapping with frequencies cut off which is determined of two filters by center frequency. Figure 8 shows graphically the Mel filters.

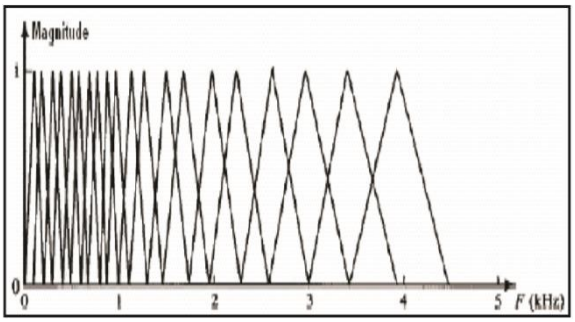


Figure 4 - Mel Filter Bank with Overlapping Filters

vii. Cepstrum

The Mel spectrum obtained, here is converted with the DCT algorithm help, back to time domain. The result we obtain is Mel Frequency Cepstral Coefficient vector values.

F- Formants and related features

Over time tracking formants is used for vocal tract shape changes modeling. Modeling formants using Linear Predictive Coding (LPC) is widely used in synthesis speech. Petrushin suggests in the prior work done by him that formants carry emotional content information. Using LPC with 15ms length frames of speech can estimate the first three formants and their bandwidths.

We calculated, for each of the three formants, their derivatives and bandwidths and across all frames the

ISSN : **2616-9185**

www.mecsj.com

variance, mean, minimum and maximum.

We also calculate the variance, mean, minimum and maximum of each formant frequency mean, its bandwidth and derivative. The vector of formant feature is 48-dimensional.

VI- **Classification**

Differentiating between "opposing" emotional states, six pairs were chosen of different "opposing" emotion:

- Despair and elation
- happy and sadness
- Interest and boredom
- Shame and pride
- Hot anger and elation
- Cold anger and sadness.

We formed data sets for each emotion pair, comprising from all speakers of emotional speech, only female speakers and only male speakers, because the gender of the speaker is affecting the features. By example, the males range pitch is from 80Hz to 200Hz while the females range pitch is from 150Hz to 350Hz. This corresponds to a total of eighteen data sets unique.

We formed inputs, for each data set, to our classification algorithm comprising of feature vectors from:

- Pitch only
- MFCCs only
- Formants only
- Pitch & MFCCs
- Pitch & Formants
- MFCCs & Formants
- Pitch, MFCCs & Formants.

Hence, the classification algorithm, for each emotion pair was running on twenty-one different inputs sets.

G- Support Vector Machines (SVMs)

The classification problem is defined in this section also it shows how presenting an approximate solution for it with the Support Vector Machine classifier.

A classification task, such as emotional state classifying, involves a training data set Strain and data Stets testing, each one containing a data instances set i ∈ S. Each instance of i is a tuple (l_i, f_i), where $l_i \in \{1, -1\}$ is a class label, with the class indicating by 1 and −1, and $f_i \in$ Rn is a feature attributes set. Producing as model with the use of Strain in the main goal of classification which predicting the labels of Stets class and gives a set of features f_i.

The Support Vector Machine (SVM) is founded in this work, that algorithm of machine learning producing the highest accuracy of emotion classification. In the SVMs case, we construct an N-dimensional hyperplane to create the model that separates optimally data into two different categories. To be the maximal between the two classes separation, optimality is taken. Any such

Multi-Knowledge ELECTRONIC COMPREHENSIVE JOURNAL FOR EDUCATION AND SCIENCE PUBLICATIONS (MECSJ)

ISSUE (13) , Oct (2018)

ISSN : **2616-9185**

www.mecsj.com

hyperplane can be written as the points x = f_i set satisfying:

$w \cdot x - b = 0$

where the normal vector is w perpendicular to the hyperplane.

We want to choose b and w such that the distance is maximized between two hyperplanes parallel separating the data between the two classes. These hyperplanes can be described with:

$W \cdot x - b = 1$

$W \cdot x - b = -1$

With the distance given by $\frac{1}{||w||}$. Thus, the distance between the two planes is maximized, we wish, while satisfying the constraints, to minimize ||w||:

$W \cdot x - b \geq 1$

$W \cdot x - b \leq -1$

Respectively For the first and second class, or:

l_i $(W \cdot x - b) \geq 1$

Since $l_i \in \{1, -1\}$. Figure 9 shown graphically these relations.

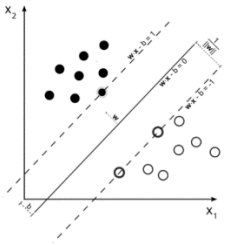


Figure 5 - Maximum separating hyperplane with margins for two classes

viii. Practically computable version

The optimal hyperplane computation reducing to solve the problem of a Lagrange multiplier optimization is shown in this section. Since the norm||w|| involves a square root, it's difficult to compute, so ||w|| is replaced by $\frac{1}{2}||W||^2$ in practice, with factor $\frac{1}{2}$ used for convenience. The solution is derived by solving the problem of Lagrange multiplier optimization:

$$\min_{w,b,\beta}\{\frac{1}{2}\left||w|\right|^2 - \sum_{i=1}^{n}\beta_i[l_i(w.x_i - b) - 1]\}$$

Where n=$|S_{train}|$ and β_i are Lagrange multipliers.

ix. Extended version – mislabeling and kernels

In this work, a SVMs extended version allowing for examples mislabeled is used. A hyperplane as cleanly as possible is chooses by this Soft Margin approach even if there is no hyperplane for the two classes. This degree of misclassification is measured by

Multi-Knowledge ELECTRONIC COMPREHENSIVE JOURNAL FOR EDUCATION AND SCIENCE PUBLICATIONS (MECSJ)

ISSUE (13) , Oct (2018)

ISSN : **2616-9185**

www.mecsj.com

the variable ξ_i and require the solution of the optimization problem:

$$\min_{w,b,\xi}\{\frac{1}{2}||W||^2 + C\sum_i \xi_i\}$$

Under constraint

l_i (w·x_i −b) ≥ 1−ξ_i 1 ≤ i ≤ n

$\xi_i \geq 0$.

Where C > 0 is the mislabeled examples penalty. Using Lagrange multipliers this can be solved as in first Equation.

Moreover, this work will use a non-linear classifier, rather than using a linear classifier. This replaces the dot product linear $x_i \cdot x_j$ by a kernel function which transforming into a higher-dimensional feature space the original input space, and allows the SVM to be non-linear and separate the two classes better potentially. After trailing several candidates kernel function possible, and includes the polynomial K (x_i, x_j) =$(x_i \cdot x_j)^d$, the Radial Basis Function (RBF).

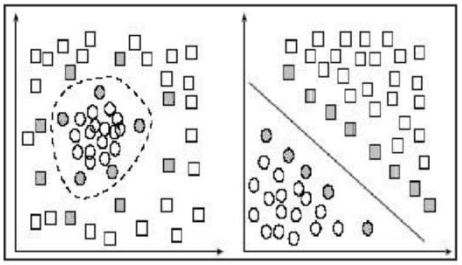
K (x_i, x_j) = exp $(-\gamma||x_i - x_j||^2)$

With coefficient of exponentiation γ > 0, was found to yield the best results promising. Though linear classification followed by SVM, it can also perform data points non-linear mapping with uses its kernel functions (Hsu, Chang & Lin, 2010).

The three kernel functions hold by SVM as following:

- Polynomial kernel
- Linear kernel
- Radial Basis Kernel.

The RBF kernel has less numerical difficulties as in Figure 10, when comparing with other kernel functions.



(a) Radial Basis Function (b) RBF mapping

Figure 6 - SVM Kernel Functions

x. Grid search

By using SVM kernel of the Radial Basis Function, and choose the penalty C for examples mislabeled and the constant of exponentiation γ in second and third equations above becomes important. Because the values optimal are model-specific, needing a search algorithm to find a set of values near-optimal.

Identifying good (C, γ) to values is the goal so that unseen data testing can be predicted by the classifier accurately, Stets, rather than maximizing the training data prediction accuracy by choosing them, Strain, which already known the labelling. Pairs of (C, γ) can be in a range

ISSN : **2616-9185**

sequentially tried to achieve this, picking the highest Stets accuracy of pair.

xi. Pairwise fusion

To generalize more than two SVMs classes, we use pairwise classification. By building for each pair of classes a classifier, a single problem of multiclass is reduced into problems of multiple binary, by using just two classes instances at a time. The two classes probabilities are then the output of the pairwise machines.

Determining the most probable class, the multiple binary SVMs probabilities is needed to be fused. There are multiple ways to fuse it into an existed ranked list, including for example:

- Majority voting, where each one of pairwise decision is counted as a class vote.
- Maximum combined probability, where the classes are ranked to their total probability according from the SVMs binary.
- Votes above a threshold, where returning the classes receiving pairwise votes over a certain threshold.

xii. Experimentation and Result

To classify all input sets for each emotion pair, a 2-class SVM classifier version modified in Schwaighofer's SVM Toolbox was used. The two kernels used with their parameters are listed below:

- Linear Kernel (with parameter C, for the coefficientsα_i's corresponding to the upper bound, with multiplicative step 10, ranges from 0.1 → 100).
- Radial Basis Function (RBF) Kernel (parameter C, for the coefficients α_i's corresponding to the upper bound, with multiplicative step $\sqrt{2}$, ranges from 0.1 → 10)
- The variance and recognition accuracy was calculated as for K-means by using the same technique. With each experiment, on achieve the highest accuracy recognition, its variance, the clustering parameters and feature inputs used is listed below in Table 2.

H- K-Means Clustering

xiii. K-Means Algorithm Overview

K-means clustering is very popular approximate method also it's an elementary that can be used to accelerate and simplify convergence. The main goal for it is to find K mean vectors (μ_1, ...,μ_k) which will be the K centroids cluster. It is traditional to chosen randomly from the data set K samples and let it to serve as initial cluster centers. The algorithm is as below:

Multi-Knowledge ELECTRONIC COMPREHENSIVE JOURNAL FOR EDUCATION AND SCIENCE PUBLICATIONS (MECSJ)

ISSUE (13) , Oct (2018)

ISSN : **2616-9185**

www.mecsj.com

- Step 0: set K the number of clusters.
- Step 1: initialize (μ_1, ...,μ_k) cluster centroids.
- Step 2: classify according to the nearest μ_k the samples.
- Step 3: precompute, until there is no significant change, (μ_1, ..., μ_k).

In figure 11 this method is illustrated. At an iteration given (top left), distributing points into two clusters. Computing their centroids (top right). Then reallocating data points to the nearest centroid cluster (bottom right) then the new cluster centroids are computed (bottom left).

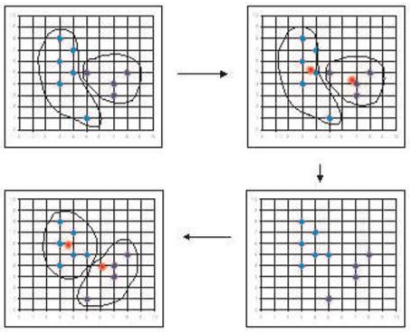


Figure 7 - Illustration of the K-means algorithm

The algorithm computational complexity is O (NdKT), where N is the samples number, d the data set dimensionality, K the clusters number and T the iterations number. Generally, a global minimum over the assignments does not achieved by K-means. In fact, since discrete assignment is used by the algorithm rather than some continuous parameters set, we cannot even properly call the minimum reaches as a local minimum. In addition, results can in a good way depend on initialization. With this method, cannot detect non-globular clusters which have a chain-like shape. Although these limitations, the algorithm is fairly frequently used as a result of its implementation ease.

xiv. Partitional clustering

Attempts to decompose directly the data set into a disjoint cluster set. More specifically, they attempt to determine a partitions valid number that optimize a certain function of criterion. The function of criterion may emphasize the local or global data structure and its optimization is a procedure iterative.

xv. Hierarchical clustering

Successively proceeds by merging smaller clusters either into larger clusters, or by splitting larger ones. The algorithm result is a dendrogram (tree) of clusters which shows the clusters relations. By cutting, at a desired level, the dendrogram into a data items clustering the disjoint groups is obtained. Because of computational time, these methods are not frequently suitable for large data sets.

Multi-Knowledge ELECTRONIC COMPREHENSIVE JOURNAL FOR EDUCATION AND SCIENCE PUBLICATIONS (MECSJ)

ISSUE (13) , Oct (2018)

ISSN : **2616-9185**

www.mecsj.com

xvi. Density-based clustering

Groups into clusters based, neighboring objects with density conditions. This model is based and parametric. Gaussian Mixture Models (or Gaussian distributions weighted sum) is a typical example.

xvii. Grid-based clustering

This method is proposed mainly for data mining spatial. the main characteristics of this method is that they quantize into a finite cells number the space and they do all their operations on the space quantized. An example of this methods is Self-Organizing Feature Maps.

xviii. Fuzzy K-means

In every classical K-means procedure iteration, each data point is assumed to be in exactly one cluster. This condition can be relaxed with assuming that each sample X_i has a number of graded or fuzzy membership in a clusterw_k. These memberships are corresponded to the probabilities p ($w_k|X_i$) that is given a sample X_i belongs to class w_k.

The clustering algorithm of fuzzy K-means seeks a minimum cost function of a heuristic global:

$$J_{fuz} = \sum_{k=1}^{K}\sum_{i=1}^{N}[p(w_k|x_{i)}]^b d_{ik}$$

Where: $d_{ik} = ||X_i - \mu_k||^2$ and b is chosen to adjust the different clusters blending as a free parameter. If b = 0, J_{fuz} is merely a sum-of-squares criterion with assigning each pattern to a single cluster. For b > 1, the criterion will allow each pattern to be belonged to multiple clusters. The solution will minimize the criterion fuzzy. In other words, one will compute the following for each 1 ≤ k ≤ K:

$$p(w_k|x_i) = \frac{(d_{ik})^{(\frac{1}{1-b})}}{\sum_{l=1}^{K}(d_{il})^{(\frac{1}{1-b})}}$$

and for each dimension 1 ≤ j ≤ d:

$$\mu_k(j) = \frac{\sum_{i=1}^{N} x_i(j)[p(w_k|x_i)]^b}{\sum_{i=1}^{N}[p(w_k|x_i)]^b}$$

The probabilities and cluster means, according to the following algorithm, are estimated iteratively:

- Step 0: set K the number of clusters and b the parameter.
- Step 1: initialize (μ_1, ...,μ_k) cluster centroids and (p ($w_k|$ X_i)) ik.
- Step 2: precompute (μ_1... μ_k) centroids and (p ($w_k|$ X_i)) ik until there is no significant change.

The probabilities incorporation as sometimes graded memberships improve the K-means convergence over its

Multi-Knowledge ELECTRONIC COMPREHENSIVE JOURNAL FOR EDUCATION AND SCIENCE PUBLICATIONS (MECSJ)

ISSUE (13) , Oct (2018)

ISSN : **2616-9185**

www.mecsj.com

counterpart classic. One drawback of the method is that the membership probability of a point in a cluster depends implicitly on the clusters number, and when this number is incorrectly specified, results can be misleading.

xix. Experimentation and Result

For each pair emotion, all input sets using K-Means clustering (k = 2) were clustered for all twelve combinations of the following parameters:

- Distance Measure Minimized: L1 norm, Squared Euclidean, Correlation and Cosine (the Correlation and Cosine distance measures is used as they defined in the MATLAB "K-Means" function).
- Initial Cluster Centroids: User Defined Centroids (UDC) and Random Centroids (RC). A UDC is the centroid used for minimizing the measure of distance of the features input for one emotion in the emotion pair.
- Maximum Number of Iterations: 1 (only when the initial cluster centroid is a UDC) and 100 (for both Random and UDC centroids).

To obtain the recognition accuracy the error used is the training errors obtained by 10-fold cross validation average, and it's a generalization error estimate. The recognition accuracy variance is these training errors variance. For each experiment, the highest accuracy of recognition achieved, its variance, the clustering parameters and inputs feature used, is listed in Table 3.

Table 2 - Highest Recognition Accuracies using K-means Clustering

VII- **Discussion**

The results obtained by the experiments performed allow us to make the following observations. Using to our classification algorithms the formant feature vector as an input, always results accurately in sub- optimal recognition. We can infer that not much emotional information is carried by formant features. Since formants are used to model the resonance frequencies (and shape) of the vocal tract, we can postulate that different emotions do not significantly affect the vocal tract shape.

Using, as a distance measure, Squared Euclidean for K-means always results accurately in sub-optimal recognition. Using this distance metric places a lot of weight effectively on an element magnitude in the feature vector. Hence, an input feature that might vary a lot between the two opposing emotions may be discounted by this distance measure, if the mean of this feature is smaller than that of other features.

Tables 2 and 3 indicate that when the classifying separately emotion pairs of male and female speakers the recognition accuracy is higher. We postulate two reasons for this behavior. First, using a big number of speakers (as in the all speaker case) increases the variability associated with the features, correct classification thereby hindering. Second, since speaker recognition is using MFCCs, we hypothesize that the features also carry information relating to the identity of the speaker. In addition to MFCCs and Pitch emotional content also carry information about the speaker gender. This additional information increases misclassification and its unrelated to emotion.

Tables 2 and 3 also suggest that the rate for female speaker's recognition is lower than male speakers when classifying emotional states of happiness, elation and interest. The higher number of female speakers than male ones in our data set may contribute to the accuracy of this lower recognition. Suggesting in further investigation that in excitable emotional states such as happiness, elation and interest, the Pitch and MFCCs variance increases significantly. However, the Pitch and MFCCs variance is higher for female voices than male voices. Hence, this variance increase is masked by the natural female voices variance, which could make the features at female speakers correctly classifying agitated emotions, less effective.

Of all the methods implemented, linear kernel SVMs gives us the best results for the classification of single-gender, in male speakers especially. This indicates that this space feature is almost separable linearly. Using K-Means classification the best results are obtained usually when UDCs are the cluster centroids which we think, unless we add some bias to the features, cannot pick up on all the information contained in the feature sets that indicating unsupervised learning algorithms such as K-Means.

VIII- **Conclusion**

Although accurately compare recognition accuracies is impossible from this study to other studies because of using a different data sets, the implemented methods here are promising extremely. The obtained recognition accuracies using SVMs for male speakers with linear kernels are higher from any other study. Previous studies have neglected to separate female and male speakers out. This project showing that there is significant benefit in doing this. Our methods are accurate at recognizing emotions reasonably in female and all speakers. Our project showing that features generated from agitated emotions such

as elation, interest and happiness have properties similar, as do those from more emotions subdued such as sadness and despair. Hence, 'subdued' and 'agitated' emotion class can encompass these emotions. This is useful especially for animating avatars gestures in virtual worlds.

2-way classification were focused within this project. These methods performances should be for multi-class classification evaluated (using multi-class SVMs and K-Means). In addition, the features could be fit to Gaussians and classified using Models of Gaussian Mixture. A numbers and dates in various emotions are uttered with the speakers used here, the words themselves carried to information emotional. In reality, emotion can be indicated by word choice. In speech recognition systems MFCCs are widely used and also carry information emotional. Existing systems for speech recognition could be modified to detect emotions as well. Improving emotion recognition could be done by combining this project methods and similar methods to the Naïve Bayes in order to take advantage of the words emotional content.

IX- **References**

[1] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002S28

[2] R.Banse, K.R.Scherer, "Acoustic profiles in vocal emotion expression", Journal of Personality and Social Psychology, Vol.70, 614-636, 1996

[3] T.Bänziger, K.R.Scherer, "The role of intonation in emotional expression", Speech Communication, Vol.46, 252-267, 2005

[4] F.Yu, E.Chang, Y.Xu, H.Shum, "Emotion detection from speech to enrich multimedia content", Lecture Notes In Computer Science, Vol.2195, 550-557, 2001

[5] D.Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)", Speech Coding & Synthesis, 1995 [6] http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html

[7] S.Kim, P.Georgiou, S.Lee, S.Narayanan. "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features", Proceedings of IEEE

ISSN : **2616-9185**

Multimedia Signal Processing Workshop, Chania, Greece, 2007

[8] http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/

[9] L.R.Rabiner and B.H.Juang. "Fundamentals of Speech Recognition", Upper Saddle River; NJ: Prentice-Hall, 1993

[10] V.A Petrushin, "Emotional Recognition in Speech Signal: Experimental Study, Development, and Application", ICSLP-2000, Vol.2, 222-225, 2000

[11] L.R.Rabiner and R.W.Schafer. "Digital processing of speech signals", Englewood Cliffs; London: Prentice-Hall, 1978 [12] http://ida.first.fraunhofer.de/~anton/software.html

[12] Julien Neel. "Cluster analysis methods for speech recognition", February 2005.