SPEECH EMOTION RECOGNITION – A TECHNICAL REVIEW OF PROCESSES

Dupinder Kaur

ABSTRACT : ***Speech emotion recognition refers to recognizing the impact and type of speech a user uses . Speech emotion is quite helpful now these days in the crime branch also as a lot of theft can be cached using this analysis . It involves primarly two section. The first section is called the training section and the other section is called the testing section . In the training section the system is trained using speech files as per their emotion and in the testing set classifiers are used for the testing purpose. This paper focuses on different analysis techniques of speech emotion and recognition system.***

KEYWORDS: ***SPEECH EMOTION , CLASSIFICATION , TRAINING***

INTRODUCTION

The dynamic requirements of automated systems have pushed the extent of recognition system to consider the precise way of command rather to run only on command templates. The idea correlates itself with the speaker identification at the same time recognizing the emotions of speaker. The acoustic processing field not only can identify „who" the speaker is but also tell „how" it is spoken to achieve the maximum natural interaction. This can also be used in the spoken dialogue system e.g. at call centre applications where the support staff can handle the conversation in a more adjusting manner if the emotion of the caller is identified earlier. The human instinct recognizes emotions by observing both psycho-visual appearances and voice. Machines may not exactly emulate this natural tendency as it is but still they are not behind to replicate this human ability if speech processing is employed. Earlier investigations on speech open the doors to exploit the acoustic properties that deal with the emotions. At the other hand the signal processing tools like MATLAB and pattern recognition researcher"s community developed the variety of algorithms (e.g. HMM, SVM) which completes needed resources to achieve the goal of recognizing emotions from speech.

This paper focuses on technical challenges that arise when equipping human-computer interface to recognize the user vocal emotions. Starting with the system

overview, the acoustic properties of voice, their features extraction and selection based on the emotion relevance which is identified by the earlier studies is reviewed and later the previous work by the speech processing community dealt with emotions is discussed in details[1].

SPEECH EMOTION RECOGNITION SYSTEM

In a generalized way, a speech emotion recognition system is an application of speech processing in which the patterns of derived speech features (MFCC, pitch) are mapped by the classifier (HMM) during the training and testing session using pattern recognition algorithms to detect the emotions from each of their corresponding patterns. The technique is synonymous to speaker recognition system but its different approach to detect emotions makes it intelligent and adds security to achieve better service in various applications.

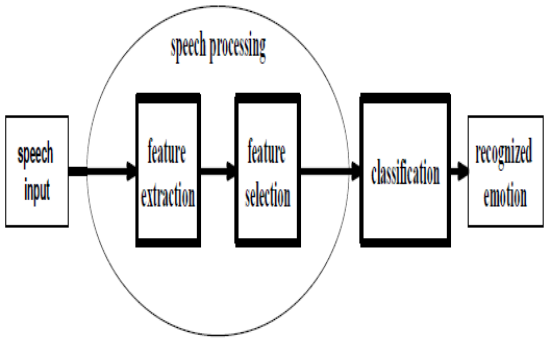

Figure. Basic Speech Emotion Recognition System

After getting the prior knowledge from previous studies we can draw the basic modular flow of the system processes as shown in Figure. Since the emotion is to be detected from the input speech signal the whole signal processing[2] revolves around the speech signal for the extraction and selection of speech features correspond to emotions. The next is generating a database for training and testing of extracted speech features followed by the last stage of emotion detection by the classifier section using pattern recognition algorithms. Firstly the speech signal is pre-processed for the removal of noise and d.c components to pass it further for features extraction and selection. The speech features are the acoustics information usually derived from the analysis of speech in both time as well as frequency domain. The extracted features are then selected in terms of emotion relevance and also to control the dimensionality of combined features which will further be classified to determine the emotion in speech as discussed in details in the further section.

Segmentation Unit

First step towards segmentation of speech signal into voiced, unvoiced and silence is Voice Activity Detection (VAD). We use

the algorithm proposed by Rabiner and Sambur for doing the VAD [5]. Firstly we calculate Zero Crossing Rate (ZCR) and Short Time Energy for each of the 30 ms long, 50 % overlapping frames obtained after applying a rectangular window function on the speech signal. Then simple energy based threshold detection is done to estimate the active part[3]. Next a smoothing algorithm is applied. Then we use the ZCR to extend the endpoints of an active area. This concludes the VAD.

The autocorrelation function of a segment of voiced speech should show higher values that the one of an unvoiced speech sound. Voiced segments generally also have low ZCR Value. So we keep a low voicing and ZCR threshold to avoid losing many voiced segments. Then we check for other segments in the active part which are not classified as voiced and have ZCR value greater than a set threshold and classify them as unvoiced. The rest of segments in active region are classified as silence/pauses. Figure 2 displays the output of our segmentation unit. For example from time interval 1.3 to 1.6 ms approx. the value of ZeroCrossing (dark blue line) is well above the ZCR threshold (magenta line) and also energy (light blue line) is present in that interval so it is classified as unvoiced[4]. Similarly from time interval 0.2 to 0.4 second energy (light blue line) and harmonicity (bold dark green line) are high whereas ZeroCrossing (dark blue line) is less than threshold so it is classified as voiced. In between interval 0.4 to 0.6 there is no energy (light blue line) so it is classified as silence/ pause.

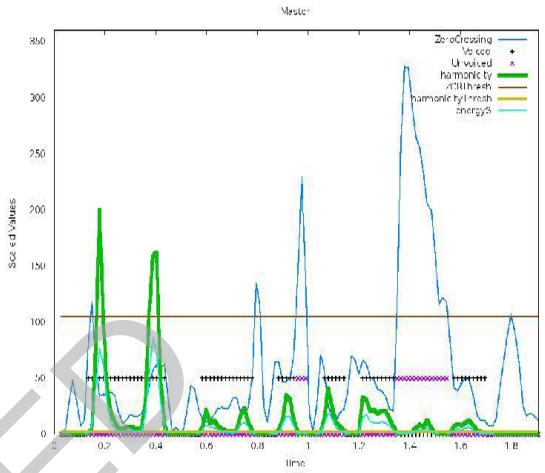

Fig. 2. Segmentation unit showing voiced and unvoiced parts.in a speech signal from EMO-DB

Feature Extraction

Extracting valuable features is another challenging task in the emotion recognition system. Mel frequency cepstral coefficients (MFCC) are one of the important features used in speech signal processing. Initially designed for speech recognition tasks they often give excellent performance in emotion detection tasks as well. We therefore calculate the average, standard deviation and minima and maxima for 17 MFCCs. Next we extract Loudness features. The Loudness is a measure for how loud or how soft a sound

is perceived-relatively independent of actual amplitude level of the signal. We apply bark filter banks [6] on the signal and then calculate loudness as described by Zwicker[4]. The center frequencies for the corresponding filters are calculated according to bark scale.

$$\Omega(k) = 6 * \log\left[\frac{f(k)}{600} + \sqrt{\left(\frac{f(k)}{600}\right)^2 + 1}\right]$$

$\text{fmin} \le f(k) \le \text{fmax}$

The shape of the filters is defined by equation:

$$Ck(w) = \begin{cases} 10^{1(\Omega - \Omega(k)) + 0.5}, \Omega \le \Omega(k) - 0.5 \\ 1, \Omega(k) - 0.5 < \Omega < \Omega(k) + 0.5 \\ 10^{-2.5(\Omega - \Omega(k)) - 0.5}, \Omega \ge \Omega(k) + 0.5 \end{cases}$$

Ck(w) is a weight of the k filter at frequency w , □(k) is a center frequency of the filter ,k = 1,2,3…..,K. Next we calculate energy and pitch related features. To calculate the pitch we use cepstrum based pitch determination. In general we observe that there is a peak in the cepstrum at the fundamental period of the input speech segment. The position of this peak should then correspond to the pitch period. We also use the short time energy features obtained during segmentation. For the energy and the loudness feature set we calculate the average, standard deviation and minima and maxima.

Next we calculate the above statistics on the first and second order derivatives of the contours in order to exploit temporal behavior at certain point of time.

Next we use the rhythm features used in linguistics including V, the standard deviation of vocalic internals (Vs); C, the standard deviation of consonantal internals (Cs); and %V, the relative duration of vocalic intervals within the total utterance etc. Our experiments seek to establish an analogy of vocalic[5] and consonants intervals to the voiced and unvoiced parts respectively. Rhythm features namely mean voiced duration, mean unvoiced duration, mean pause duration, standard deviation of voiced, unvoiced, and silence intervals are calculated. VarcoV feature given by Eq.1 is then calculated for voiced, unvoiced and silence intervals.

VarcoV □ std (V) *100 / mean(V)

Similarly we calculate number of voiced intervals per sec and the rhythm feature nPVI for voiced, unvoiced and
silence regions which is given by Eq.2 as:

$$nPVI = 100 * \sum_{k=1}^{m-1} \left| \frac{V(k) - V(k-1)}{(V(k) + V(k+1))/2} \right| / (m-1)$$

Next we calculate temporal features such as:

- ✓ duration of pause / duration of voiced+unvoiced
- ✓ duration of voiced/ duration of unvoiced

- ✓ duration of unvoiced / duration of unvoiced+ voiced
- ✓ duration of voiced / duration of unvoiced+ voiced
- ✓ duration of voiced/ pause(silence)
- ✓ duration of unvoiced / pause(silence)

In whole we get a set of 487 features. Table 1 shows a summary of extracted and calculated features and the number of features respectively[7].

Table 1. Extracted Features

Feature Source	Number of Features
MFCC's voiced	204
MFCC's unvoiced	204
Loudness Voiced	12
Loudness Unvoiced	12
Pitch	12
Energy	24
Rhythm and Temporal	19

Feature Selection

In order to determine the most promising features for our task individually, we applied an Information Gain Ratio (IGR) filter. We use WEKA toolkit .This entropy-based filter estimates the goodness of a single attribute by evaluating its information contribution (gain) with respect to the required mean information that leads statistically to a successful classification. The final ranking is obtained by using 10 fold cross validation. Table 2 shows the top 20 ranked features.

To select optimized number of features for classification we expand the feature space by including more and more features starting with the high ranks of the IGR output as shown in Figure. We observe that selecting 305 features gives the best recognition[8]. As expected, the recognition rates rise by adding features until a global optimum is reached. Including more features beyond this optimal number causes a degradation of recognition rates again, due to adding irrelevant or even harmful information to the classification.

Classification

In our Classification experiments compare the classification results achieved by using different features separately as well as in combinations. We evaluate the performance of two different classifiers as shown in Table 3. The first classifier ANN is implemented following the multilayer perceptron architecture, using WEKA software. An artificial neural network (ANN), usually called "neural network" (NN), is a mathematical model or computational model that tries to simulate the structure and/or functional aspects of biological neural networks. It consists of

an interconnected group of artificial neurons and processes information using a connectionist approach to computation [9]. After experimenting with different network parameters highest accuracy is found by using 200 neurons in hidden layer. The learning and momentum rate are left to the default setting of WEKA (0.3 and 0.2 respectively). The number of epochs is set to 500. Error backpropagation is used as a training algorithm. As a second classifier we choose a Support Vector Machine (SVM). John Platt's sequential minimal optimization algorithm is used for the optimizing and a Polynomial kernel of first order is used [10]. The value of cost parameter is kept to be 1. Now we do speaker dependent experiments using these classifiers.

Table 2. Information Gain Ranking

Rank	Feature
1	mfcc_max_org_1_voiced
2	mfcc_mean_org_1_voiced
3	mfcc_mean_org_16_voiced
4	mfcc_mean_org_2_voiced
5	mfcc_min_org_1_unvoiced
6	mfcc_min_org_1_voiced
7	pitch_min_org_voiced
8	pitch_mean_org_voiced
9	mfcc_max_org_2_voiced
10	mfcc_mean_org_15_voice
11	mfcc_mean_org_17_voiced
12	mfcc_max_org_15_voiced
13	mfcc_max_org_14_voiced
14	mfcc_mean_org_14_voice
15	mfcc_max_org_13_voiced
16	mfcc_mean_org_5_unvoiced
17	mfcc_mean_org_6_unvoiced
18	rhythm_mean_voiced
19	mfcc_min_org_2_unvoiced
20	mfcc_max_org_10_voiced

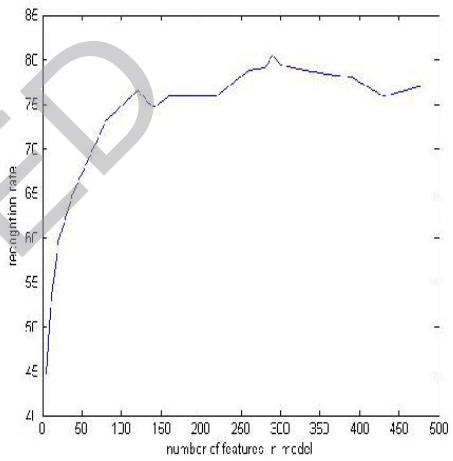

Fig.3. Recognition rate versus number of selected features from IGR ranking

We display results in form of confusion matrix and % recognition. After selecting the top 305 features by the selection algorithm described in previous section we are able to achieve recognition rate of 80.60 % on Berlin Emotion Database for 7 emotions classification problem. For the classification experiment the 10x10-fold stratified cross-validation method is employed over the data sets.

Table 3. Recognition rates for different classifiers

Classifier	% Recognition
SVM	80.27 %
ANN	80.60%

Table 4 and Table 5 show the confusion matrices for SVM and ANN. From Table 6 we interpret that Sadness is the best recognized emotion with accuracy of 95.2 %, 88.71 % and Happiness is worst recognized with accuracy of 66.2 %, 60.56 % for ANN and SVM classifier respectively. SVM classifier gives better accuracy for Boredom and Neutral Classes in comparison to ANN which gives better accuracy for Happiness, Sadness and Disgust Classes.

Table 4. Confusion matrix after feature selection (SVM Classifier)

B	N	A	F	H	S	D	EMOTION
66	9	0	0	0	2	2	B (Boredom)
5	67	0	4	2	1	0	N (Neutral)
0	1	106	2	14	0	3	A (Anger)
0	4	5	55	2	2	0	F (Fear)
2	2	18	5	43	0	1	H (Happiness)
2	3	0	1	0	55	1	S (Sadness)
3	2	2	2	1	2	34	D (Disgust)

Table 5. Confusion matrix after feature selection (MultilayerPerceptron)

B	N	A	F	H	S	D	EMOTION
61	10	0	0	0	4	4	B (Boredom)
10	62	1	2	1	2	1	N (Neutral)
0	0	106	3	14	0	3	A (Anger)
0	5	4	54	4	1	0	F (Fear)
0	2	16	5	47	0	1	H (Happiness)
1	0	0	1	0	59	1	S (Sadness)
1	0	2	3	1	0	39	D (Disgust)

Table 6. Class wise recognition rate for ANN and SVM classifier

Emotion	SVM	ANN
Boredom	83.54 %	77.2 %
Neutral	84.81 %	78.48 %
Anger	84.13 %	84.13 %
Fear	80.88 %	79.42 %
Happiness	60.56 %	66.20 %
Sadness	88.71 %	95.2 %
Disgust	73.91 %	84.7 %

Now we visualize the results obtained by different sets of features individually and in groups as shown in Table 7. We see that MFCC features alone are the best features giving a recognition rate of 71.93 %. Adding rhythm features to MFCC improves the accuracy to 74.02 % while the Rhythm features by themselves lead to only 34.6 %. However, as there are more than 10 times more MFCC features compared to the number of Rhythm features this result is expected to be influenced by quantity of features also. Loudness Features lead to only 44 % accuracy. If MFCC features are excluded then accuracy of 62.5 % is achieved. MFCC voiced (62 %) have better accuracy then MFCC unvoiced (49.3 %).

Table 7. Recognition rates for Different Sets of Features (SVM classifier)

Features	Recognition Rate (in %)
MFCC only	71.9
Loudness + Rhythm	52.6
Loudness	43.9
Except MFCC	62.5
Rhythm only	34.6
MFCC+ Rhythm	74.02
MFCC unvoiced	49.3
MFCC voiced	67.3

Next we break down the seven-class experiment into binary classification targets and get the results as shown in Table 8. Here we interpret that anger and happiness get confused very often because many of the features tend to show similar behavior for these two classes [13]. Best classification is achieved for happiness and sadness pair where accuracy of 100 percent is achieved.

Next we cluster the emotion classes into High Arousal (Happiness, Anger and Fear) and Low Arousal (Boredom,

Sadness, Disgust, Neutral) and do the analysis. We interpret from the results in Table 9; Rhythm features all alone are able to give an accuracy of 74%, MFCC alone are able to give an accuracy of 89 %, MFCC and rhythm combined result in 92.4 % accuracy and an Overall accuracy of 94 % is achieved by selecting top 305 features by the same[11] method as described in previous section.

Table 8. Recognition rates for binary classification targets (SVM classifier)

Emotion	Recognition Rate (in %)
Anger vs. Happiness	79.23
Anger vs. Disgust	95.56
Anger vs. Fear	92.04
Anger vs. Boredom	98.93
Anger vs. Sadness	99.41
Anger vs. Neutral	98.94
Fear vs. Happiness	92.5
Fear vs. Disgust	79.38
Fear vs. Boredom	98.52
Fear vs. Neutral	92
Fear vs. Sadness	95.83
Disgust vs. Happiness	92.92
Disgust vs. Boredom	92.37
Disgust vs. Neutral	96.63
Disgust vs. Sadness	95.098
Happiness vs. Boredom	99.307
Happiness vs. Neutral	93.75
Happiness vs. Sadness	100
Boredom vs. Neutral	92.42
Boredom vs. Sadness	96.79
Sadness vs. Neutral	98.94

Table 9. Recognition rates for High and Low arousal classes.

Feature	Recognition rate
Rhythm	74 %
MFCC	89%
MFCC + Rhythm	92.4%

Database for training and testing

A good database is as important as the desired results. There are different databases created by speech processing community with the help of professional actors which is widely used in research work. The results are uncompromising though the emotions are acted rather than spontaneous or natural. The famous databases are The Danish Emotional Speech Database (DES), and The Berlin Emotional Speech Database (BES), as well as The Speech under Simulated and Actual Stress (SUSAS) Database. DES and BES are representative for the early databases in the nineties but still serve as exemplars for acted emotional databases. For English, there is the 2002 Emotional Prosody Speech and Transcripts acted database available.

Classifiers to detect emotions

The selected feature vectors are stored in the database and later fed to classifier to detect the emotions by comparing the vectors from the trained data and test data vectors. There are various classifiers available meant for their specific usage based on types of features to be classified. If feature vectors belong to the global statistics SVM (Support Vector Machine), Neural Networks, Decision trees are employed and for the vectors of short-term features HMM (Hidden Markov Model) is used for its dynamic performance.

REFERENCE

[1]. Björn Schuller," Automatic Emotion Recognition by the Speech Signal",

Institute for Human-Machine-Communication, Technical University of Munich 80290 Munich, Germany, {schuller,lang,rigoll}@ei.tum.de.

[2]. Reda Elbarougy," Speech Emotion Recognition System Based on a Dimensional Approach Using a Three-Layered", Japan Advanced Institute of Science and Technology (JAIST), Japan.

[3]. Thurid Vogt, " Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realisation", Multimedia Concepts and Applications vogt,andre,wagner}@informatik.uni augsburg.de.

[4]. Carlos Busso," Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information", Emotion Research Group, Speech Analysis and Interpretation Lab, Los Angeles http://sail.usc.edu.

[5]. Dimitrios Ververidis," Emotional speech recognition: Resources, features, and methods", Artificial Intelligence and Information Analysis Laboratory, Department of Informatics, Aristotle University of Thessaloniki, Box 451, Thessaloniki 541 24, Greece.

[6]. Caglar Oflazoglu," Recognizing emotion from Turkish speech using acoustic features", Oflazoglu and Yildirim EURASIP Journal on Audio, Speech, andMusic Processing 2013, 2013:26 http://asmp.eurasipjournals.com/content/2013/1/26.

[7] O. Pierre-Yves, "The production and recognition of emotions in speech: features and algorithms," International Journal of Human-Computer Studies, vol. 59, pp. 157–183, July 2003.

[8] C.M. Lee, and S. Narayanan, "Toward Detecting Emotions in Spoken Dialogs," IEEE Transactions on Speech and Audio Processing, vol. 13(2), pp. 293–303, 2005.

[9] I. Albrecht, and M. Schroder, and J. Haber, and H.-P. Seidel, "Mixed feelings: Expression of non-basic emotions in a muscle-based talking head," Virtual Reality, vol. 8(4), pp. 201-212, 2005.

[10] D. Wu, and T.D. Parsons, and S. Narayanan, "Acoustic Feature Analysis in Speech Emotion Primitives Estimation," Proc. InterSpeech 2010, pp. 785–788, 2010.

[11] M. Schroder, and R. Cowie, and E.D.-cowie, M. Westerdijk, and S. Gielen, "Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis," Proc. Eurospeech 2001, pp. 87–90, 2001.