ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

Emotion Recognition in Speech using Inter-Sentence Time-Domain Statistics

Alexander I. Iliev

Assistant Professor, Department of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria

ABSTRACT: The main goal of this work is to show how a set of several time-domain features, as extracted from speech, can contribute for the Emotion Recognition in Speechproblem (ERS). Moreover, their fluctuations among a collection of spoken voiced events, which form multiple spoken utterances baring five distinct emotions are studied and displayed for analysis. In more details, the study compares the rate of change in time and uses their speaker independent inter-sentence derivatives while comparing its performance to the classical featureapproach. Different classification methods were employed such as Naive Bayes, Support Vectors and k-NN. In the spoken database there were clearly five distinctive emotional classes namely: *happy (H), angry (A), sad (S), neutral (N)* and *fear (F),* and for short this modelis referred to as HASNF. The system shows a very successful recognition rate of 91.67% when using the chosen time-domain feature vector. In contrast, the rate of change of these parameters is not as promising, and can be concluded that change of attributes is not as emotion-dependent as their parameters. The performance of the second system topped 44.17% correct recognition rate, which is not successful for the five-classtime-domain feature case considered here (HASNF).

KEYWORDS:Emotion, Speech, Recognition, Emotion recognition in speech, Time-domain features from speech.

I. INTRODUCTION

Emotion recognition comes in different layers and can be conveyed through behavioral patterns of different nature. They may be expressed through facial expression [1], physiological changes such as change of skin color, sweating, body movements or through speech [2, 3, 4]. The latter is of great importance because of the simple ways it can be recorded, stored, processed and transmitter. Emotion recognition from speech is a growing field in the world of human-machine interaction. With all the smart devices that we are constantly surrounded, now more than ever, we need systems to take decisions for us not only by key strokes pressed by the user, but also by reaching at the next level of what “smart” may mean. It is certainly a necessity for higher level of automation, in which based on our mood a smart device may take a decision or make a suggestion for us. It is useful in customer care, security, robotics, speech synthesis and plethora of new services in many different fields.

For the emotion recognition task from speech in this work each emotion was extracted from a series of short and continuous voiced segments of speech (VS). Several voiced regions in a sequence construct an utterance (UT). The utterances from each emotion were processed in order to find all voiced regions. All features were then extracted from all VS in every UT for each of the five HASNF emotions. Feature extraction is explained in details in chapter 2.

II. RELATED WORK

Speech has several natural domains, by means of which a multi-layered message is transmitted from person to person. Emotion is one of these layers and it can be conveyed through phonetic, prosodic, and linguistic sub-levels. It follows that acoustical analysis of speech is an important medium for the extraction of relevant signal features thatdefine variety of emotional states. One of the keyadvantages of analyzing speech when compared to other techniques is its non-intrusiveness. The fact that it can be recorded with a simple microphone over a telephone line makes it really attractive.Depending on the emotional state, a speaker changes their verbal message so that it gives the receiver enough cluesfor proper emotional perception. Some of the important parameters may befundamental pitch, energy, and speech

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

rate. It is important to target these parameters individually in any state of given speech production in orderto determine specific emotion with computational means.

When it comes to determining the finite number of emotionsof interest to be studied,the literature presents no clear consensus on the matter. Ververidis and Kotropoulos [5]presented a total of 29 different emotions that have been collected through 64 emotional speech data collections. The number of most emotional cases used in those sources varied from four to six on average. In this way, they formed a basic set, whereas other supra-level emotions were less broadly used.In other worksperformed by Cowie and Cornelius [6] and by Cornelius [7], the number of emotions is also different, however the main emotional setis formed by the so-called "Big-six": *anger*, *happiness*, *sadness*, *fear*, *surprise*, and *disgust*. This same emotional set has been used by Bhatti et al. [8], in which work it was stressed that prosody is a significant emotional medium. Other authors, such as Noda et al. [9],worked with the four emotional states: *happy*, *angry*, *sad*, and *neutral*. All of this shows that a different number of emotions can be studied for the emotion recognition problem in speech, but in order to be fair, the specifics of the particular problem have to be considered.It can be seen from the literature that most studies use from three [9] to six emotions [6]. This is why in this work some of the most basic emotions were chosen, namely *happy (H)*, *angry (A)*, *sad (S)*, *neutral (N)* and *fear (F)*.

Focusing on the feature sets and classification methods used, one can find out that there is a great variety as well. Selecting the most effective feature in a complicated classification problem such as this is a key factor for success. Some scientists focus solely on extracting features through speech signal processing [10], while others used facial recognition methodsin combination with speech signal analysis [11]. Specifically, the feature sets vary from using intonation information using the Tonal and Break Indices (ToBI) tone tier features as seen in the work done by Beckman and Elam[12], to the work performed by Gobl and Chasaide [13],in which they used the correlation between voice overtones and pitch dynamics combined with vocal tract features. Energy in speech, as well as pitch and duration have also been used in a separate study by Iida et al. [14].Other features include the use of formants [5];glottal features and specifically glottal symmetry, representing the ratio of closing to opening phase of glottal impulse [3]; Mel-Frequency Cepstral Coefficients (MFCC) of various orders [15]. When is comes to the number of types of classifiers used in the ERS problem, there is also a great variety. Some of the classifiers used in the field are: the Gaussian mixture model (GMM), Bayesian classifier (BC), k-Nearest Neighbor (k-NN), Support vector machine (SVM), C4.5 classifier, and Optimum path forest search (OPF) [15].

III. FEATURE DOMAIN

The features used were all extracted in time-domain, which is both convenient and practical. It is convenient because time-domain features require less calculations for extraction and no conversion to frequency-domain is necessary, which would have otherwise added some complexity in the overall calculation process due to the multiple additions and multiplications when using the Fast Fourier Transform. It is practical, because in an effort to create a specific solution this methodology is faster when extracting and classifying speech features in near-real time.

Feature domains can vary greatly. Many studies have used features like: mel-cepstral coefficients MFCCs [16], Tonal and Break Indices - ToBI for American English intonation [17], glottal signal extracted from free speech or recorded via laryngograph [18], pitch tracking information in time-domain and signal energy statistics also obtain in time-domain. Although each of these methods has it's pros and cons, the latter two are attractive in certain cases when dealing with a specific emotion recognition task because of relatively simple computational complexity. In these terms, the results shown here have the goal to examine their appropriate usability in a semi-supervised environment. It is such, because the training and testing stages used samples from the same speech pool with various split as mentioned later.

The feature vectors used for training and testing had fixed lengths of 11 elements:6 pitch, 3 temporal energy and 2 durational features as shown in Table 1.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

Time-domainFeatures			
Pitch elements– features & their 1st derivatives			
1	Mean	[meanP]	[dmeanP]
2	Median	[medP]	[dmedP]
3	Standard deviation	[stdP]	[dstdP]
4	Maximum	[maxP]	[dmaxP]
5	Rising-falling	[rifaP]	[drifoP]
6	Max falling range	[maxfrP]	[dmaxfrP]
Temporal Energy vs. their 1st derivatives			
7	Mean	[meanE]	[dmeanE]
8	Standard deviation	[stdE]	[dstdE]
9	Maximum	[maxE]	[dmaxE]
	Durational features vs. their 1st derivatives		
10	Zero crossing rate	[zcRate]	[dzcRate]
11	Speaking rate	[spRate]	[dspRate]

Table 1: Time-domain features.

These features can be represented as:

$$V_F = [V_p V_e V_d] , \quad (1)$$

and the rate of change is expressed as:

$$V'_F = \left[\frac{dV_p dV_e dV_d}{dt}\right], \quad (2)$$

where,

V_p = (*meanP, medP, stdP, maxP, rifaP, maxfrP*) ,
V_e = (*meanE, stdE, maxE*) ,
V_d = (*zcRate, spRate*)
dV_p , dV_e , dV_d - represent the rate of change in time respectively

and

V_F represents the complete feature vector holding all time-domain features, V_p are the pitch elements, $V_e$are the temporal energy features and V_d is the feature vector containing the two duration features. V_p , V_e and V_d were extracted from the 100 min continuous speech corpus for all five emotions HASNF and are the average of all VS for each UT. The expressions for the first derivatives of all attributes cross utterance signify the change of each feature among UTs. Since all features were very different in values from one another, and in order to avoid NaN - 'not-a-number'occurrences due to division by zero, all feature vectors were normalized using a DC-shift, which is essence introduced a linear change equally for all classes.

IV. SPEECH CORPUS

The speech corpus was designed using a theatrical play and consisted of four male speakers. Since acted speech is more exaggerated and prominent, the label markings were very accurate as there was little or no disagreement in the process. The play was 100 minutes long, with an average of 23.71 voiced events per minute and sampled at 8kHz and 16-bit resolution. The speech was transcribed in 5 distinctive emotional classes and all others were omitted. Each of the classes was represented by a different number of emotional occurrences shown in Table 2.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

Emotion:	Total # of UTs:
Happy	303
Angry	403
Sad	319
Neutral	1,215
Fear	131
All	2,371

Table 2. Emotions and their corresponding number of UTs.

As can be seen from the table, naturally the Neutral class (emotionless state) has most number of occurrences. It consisted of 2,371 spoken emotional events in total excluding long speaking pauses, noises, silence, sighs, clicks, etc. A spoken event is usually a sentence, which may be long or short. It can also consist of a single word with a different length. A simple summary of the average length of each spoken emotional event is displayed in Table 3.

Emotion:	Average UT length [sec]:	Average # of VS per UT:
Happy	9.22	4.53
Angry	4.81	4.02
Sad	11.16	3.84
Neutral	15.29	3.67
Fear	3.91	2.71
All	8.88	3.75

Table 3.Summary of spoken emotional events.

From Table 3 it follows that the total number of VS used in this emotional speaker database was 8,891. Some of the voiced events however were extremely short to be considered for the task and were eliminated. Because of that Fear fell to 122 voiced segments and in order to have a training-testing model that has equal amount of samples from each emotional class, all emotions were limited to 120. The difference from 122 is because of testing the rate of change or finding the first derivative. Therefore the final amount of all test-training utterances was 5x120 or 600 utterances. Having this in mind and considering expression (1) and (2), we can then express each emotional class (HASNF) in a matrix form for both cases as follows:

$$V_F^{HASNF} = \begin{bmatrix} V_{p1} & V_{e1} & V_{d1} \\ \vdots & \vdots & \vdots \\ V_{p120} & V_{e120} & V_{d120} \end{bmatrix}, \qquad (3)$$

and

$$V_F^{HASNF'} = \begin{bmatrix} dV_{p1} & dV_{e1} & dV_{d1} \\ \vdots & \vdots & \vdots \\ dV_{p120} & dV_{e120} & dV_{d120} \end{bmatrix}, \qquad (4)$$

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

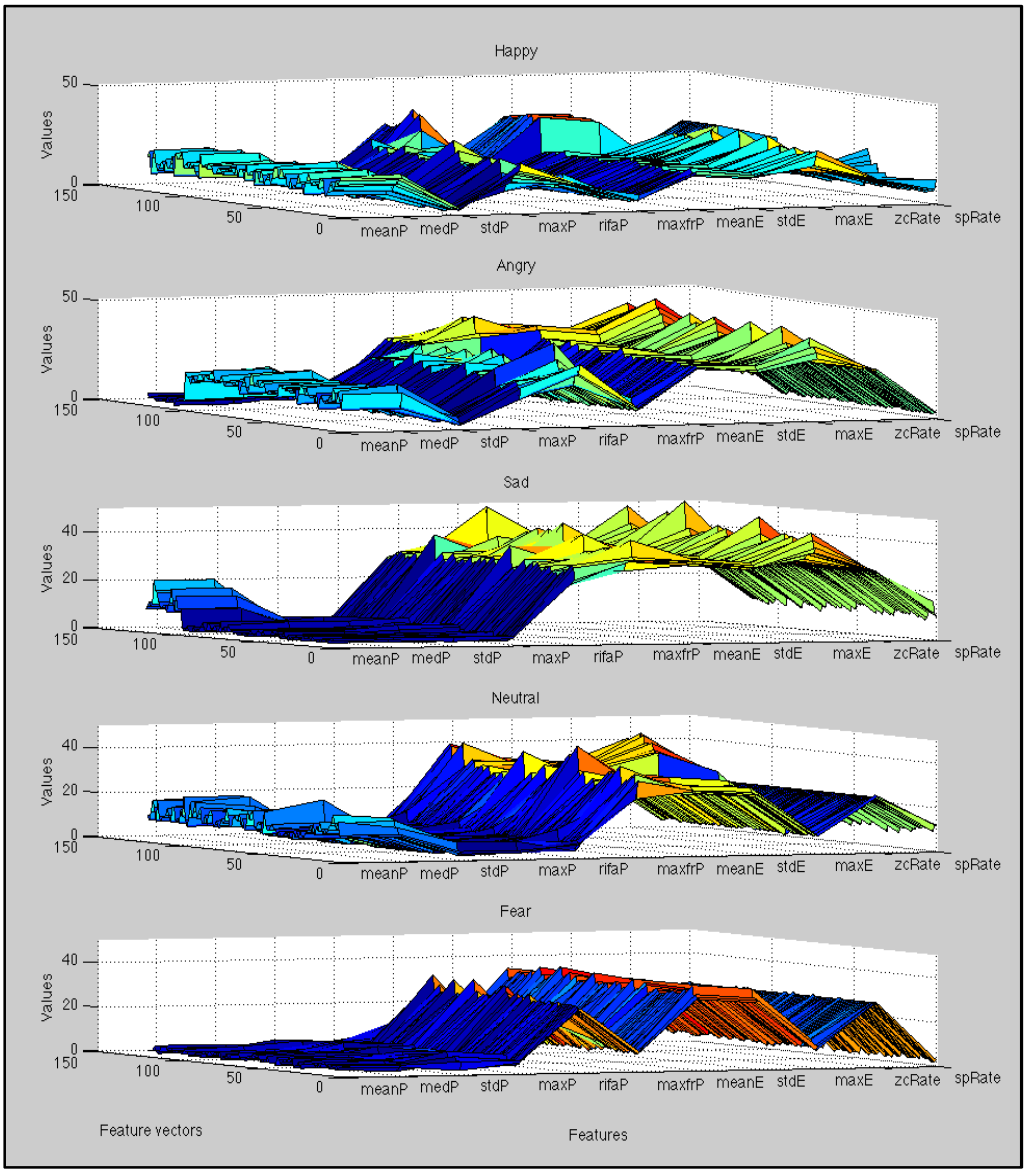

Figure 1: 3-D view of all feature vectors for each emotion.

In Figure 1 we have a 3-D display of all feature vectors for each emotion. There are 120 feature vectors, each containing the 11 features described in Table 1, and all of them with their corresponding values. It can be noted that for the most part there are visual differences in each model such as the lower first part of each feature vector, especially for the emotional cases in *sad*, *neutral* and *fear*.More detailed analysis is given in *Discussion of Results* later in this paper.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

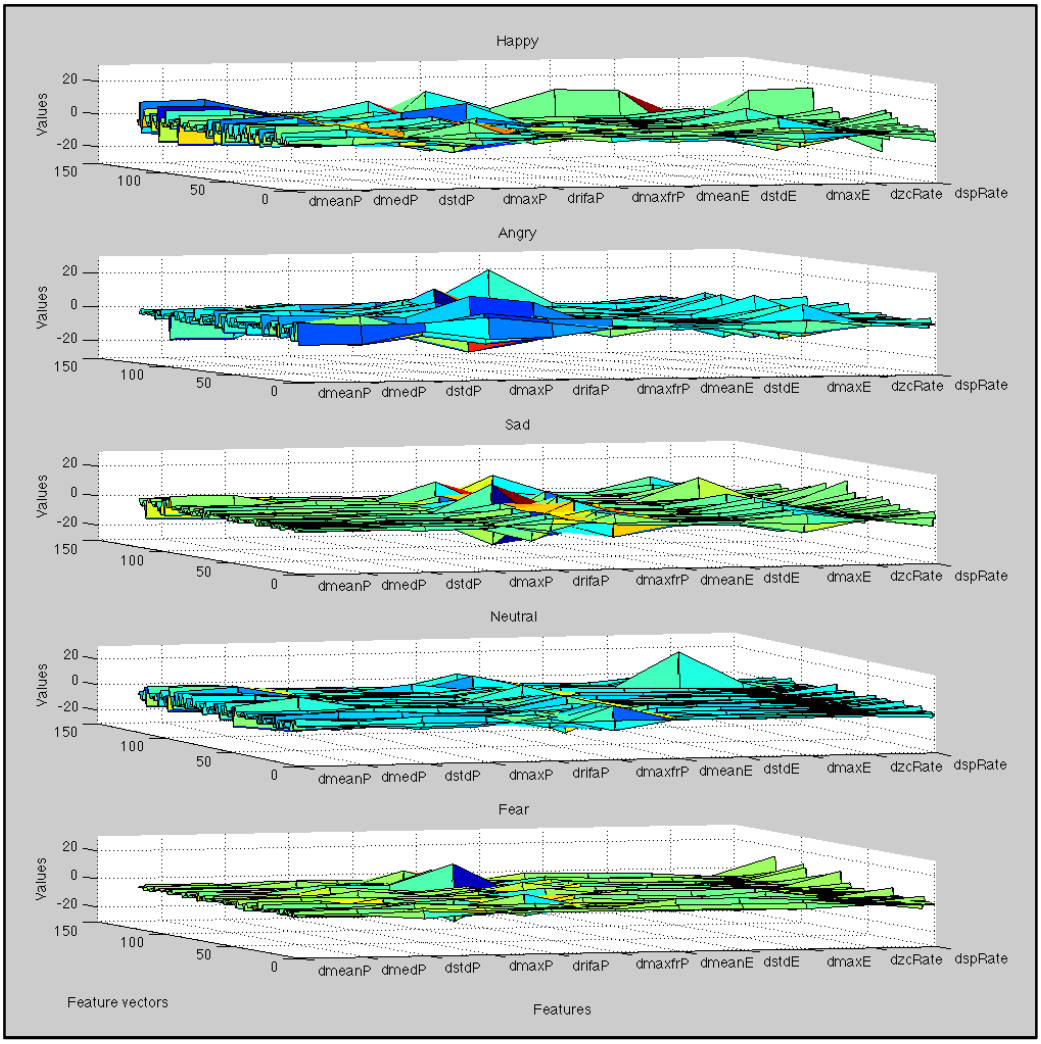

Figure 2: 3-D view of the rate of attribute change for HASNF inter-sentence.

In Figure 2 the corresponding first derivatives of feature for each emotion is shown. As a contrast to Figure 1, in Figure 2, 121 feature vectors were considered to obtain the rate of change. In this case, the through visual inspection it is obvious that the feature patterns are not as distinctive as they are from the features themselves as displayed in Fig. 1. We can argue that there is a bit more similarity between *happy* and *angry* or between *sad* and *neutral*.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

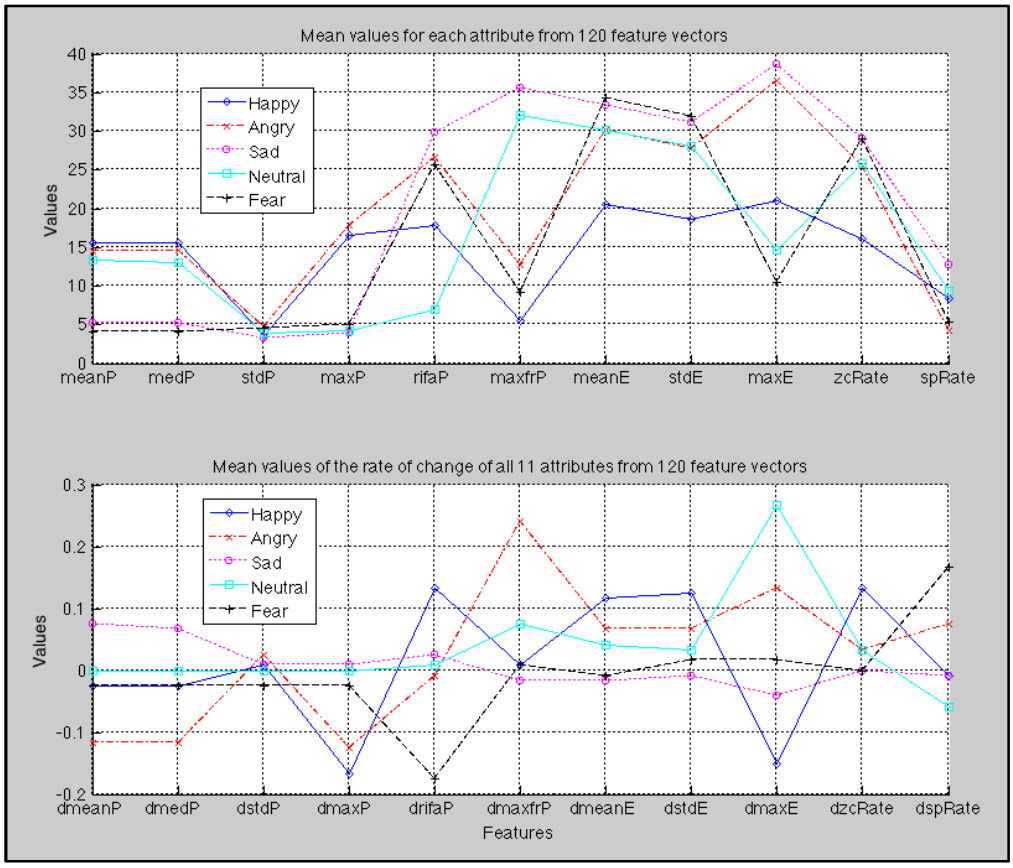

Figure 3: Mean attribute values vs. their mean rate of change.

Figure 3 shows the mean value for each attribute. In particular, the upper graph depicts the actual raw feature vectors and the lower graph displays the rate of change of these features in time. From a first glance it looks like the lower graph is more chaotic and disorganized the one above.

V. CLASSIFICATION

Several data sets were obtained for training and testing in the classification process. Since there are several voiced sections in an utterance, we need to look at the overall emotional message from an utterance rather than choosing a temporary emotional pattern. This was confirmed by testing voiced sections taken from the beginning, at the end, in the middle or at random in a given utterance. For that reason all feature vectors were constructed based on the average of all VS in a UT. This way emotion was captured on a per utterance basis and was looked at holistically. The following five classifiers were used in this study: Naive Bayes, k-NN, Support Vectors Random Forest of random trees and a Tree with randomly chosen K class attributes as implemented in Weka [19].

The Naive Bayes classifier works with independent assumptions between predictors and is an excellent tool to use for larger datasets. It did outperform all the rest of the classifiers used and mentioned earlier in this study and therefore is the only one chosen to be displayed here. In essence it calculates the posterior probability $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$, where (c) is a given class from the HASNF set subject to this study and (x) is a predictor. The effect of predictor (x) on a particular class (c) is assumed to be independent from the assessment of all other predictors. It can be expressed as follows:

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}, \qquad (5)$$

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

where,

$P(c|x)$ represents the posterior probability of one of the HASNF classes while given the predictor or attribute in V_F;

$P(x|c)$ represents the likelihood of a probability of certain predictor when given a particular class;
$P(c)$ is theprior probability of a particular class of the five HASNF;
$P(x)$ is the prior probability of certain predictor.

Naive Bayes classification using 11 time-domain features					
act \ det	H	A	S	N	F
H	28	0	0	0	0
A	2	13	8	1	0
S	0	0	25	0	0
N	0	0	0	19	0
F	0	0	0	0	25

Correctly Classified Instances	110	91.6667 %
Incorrectly Classified Instances	10	8.3333 %
Kappa statistic	0.8952	-
Mean absolute error	0.0318	-
Root mean squared error	0.1751	-
Relative absolute error	9.9138%	-
Root relative squared error	43.7357 %	-
Coverage of cases (0.95 level)	92.5000 %	-
Mean rel. region size (0.95 level)	20.1667 %	-
Total Number of Instances	120	-

Table 4: Confusion matrix for HASNF using all 11 time-domain features.

The percentage split between training/testing samples varied between 60/40, 70/30 and 80/20. The results in all cases were comparable; therefore for convenience only the 80/20 case will be provided herein. In Tables 4 and 5 are shown the results from the experimental data after being tested with Naive Bayes classifier.

Naive Bayes classification using 1st derivative of 11 time-domain features					
act \ det	H	A	S	N	F
H	7	5	4	3	9
A	1	8	3	3	8
S	1	1	9	2	12
N	0	1	1	10	7
F	2	1	1	2	19

Correctly Classified Instances	53	44.1667 %
Incorrectly Classified Instances	67	55.8333 %
Kappa statistic	0.3020	-
Mean absolute error	0.2489	-
Root mean squared error	0.3826	-
Relative absolute error	77.6994 %	-
Root relative squared error	95.5554 %	-
Coverage of cases (0.95 level)	95.0000 %	-
Mean rel. region size (0.95 level)	70.1667 %	-
Total Number of Instances	120	-

Table 5: Confusion matrix for HASNF using 1st derivative of all 11 time-domain features.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

As can be seen the 11 chosen time-domain features also known as classical approach features [20], show very high performance. The same cannot be said about the 1st derivatives for all attributesspeaker independent inter sentence case. Having the latter in mind can be concluded that the changes of all features are not as emotionally dependent as the features themselves. The 11 time-domain classical approach features were further evaluated for their gain ratio and the results are shown in Table 6:

	Feature ranking		
#	Ranking:	Feature:	V_F #
1	0.6266	zcRate	10
2	0.5792	maxE	9
3	0.5624	meanE	7
4	0.5359	rifaP	5
5	0.5344	maxfrP	6
6	0.4716	stdE	8
7	0.4573	maxP	4
8	0.4278	medP	2
9	0.4273	meanP	1
10	0.3009	spRate	11
11	0.0732	stdP	3

Table 6: Supervised gain ratio feature evaluation.

The order of importance in feature domain is top to bottom, or most to least important respectively. This order can be looked at as significance order of each feature and is somewhat consistent with the mean values for all 11 time-domain features displayed in Figure 3. In essence, Table 6 shows how much each attribute is worth by measuring the gain ratio with respect to the class.

VI. DISCUSSION OF RESULTS

As can be seen from the results, the features chosen for classifying each emotion are distinctive enough, which leads to over 90%successful emotional class recognition. We also observe that the rate of change or their first derivatives on an inter sentence level are not as prominent since classification on these derivatives is not as distinctive as the feature vectors themselves. That is to say that feature evolution for each emotion is more comparable, then the feature variations themselves among all emotional classes of interest (HASNF). In support of that notion and observing Figure 1, we see that the matrix of parameters for each emotion is somewhat visually different. We see that $V_p{}^F$ for example is much flatter than $V_p{}^{HAN}$ and only lightly resembles $V_p{}^S$. Although $V_p{}^H$ looks slightly similar to $V_p{}^A$ the two emotional classes *H* and *A* certainly differ at$V_p{}^{HA}$*[maxfrP, meanE, stdE]*and also at the durational attributesV_d. In general, all pitch related values are much higher in $V_p{}^{HA}$ than in any other class $V_p{}^{SNF}$. In fact the first four pitch parameters $V_p{}^S$*[meanP, medP, stdP, maxP]* in emotional class *F* have the flattest shapes of all the rest of the emotional classes, followed by emotional class *S*. On the other hand $V_p{}^N$*[maxE]* for emotional class *N* has a unique shape when compared to $V_p{}^{HAS}$*[maxE]*. Furthermore, emotional class *F*has a unique temporal shape$V_e{}^F$, which makes itmore distinctive than the rest of the classes. Moving on to Figure 2, from a first glance there are only small differences in attribute change rate among different classes, which also explains the unsuccessful recognition rate of the latter when compared to the attribute vectors displayed in Figure 1. Figure 3 displays the mean of each attribute value vs. its rate of change for the speaker independent inter sentence case. Furthermore, the recorded classification rate is displayed in Tables 4 and 5 clearly show that using simple time-domain features is definitely better than using their rate of change in a speaker independent inter sentence case.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

VII. CONCLUSIONS AND NEXT STEPS

This study shows how time-domain features can perform very well in certain speaker independent applications. The performance of the attribute vectors is much greater and surpassesgreatly the one of their respective rate of change in time. Regardless of these promising findings,a clear parametric distinction in the construction of a tangible ESR system with fewer computations, requires more testing on various speech corpora containing more diverse speaker pool including people from different ages and gender. Since this study is solely performed on American English speech, it would be interesting to conduct a similar study on speech corpora with different English source with accents or even an entirely different language. Emotions are to some extent dependent on culture [21, 22] and are conveyed in different ways through different nations [23]. It isinstrumental to study such case and draw an emotional speech dependent parallel between the well-studied American English field and other English accents or languages. Only then we can get a step closer to more practical applications.

REFERENCES

[1] Spiros V. Ioannou, Amaryllis T. Raouzaiou, Vasilis A. Tzouvaras, Theofilos P. Mailis, Kostas C. Karpouzis, Stefanos D. Kollias, "Emotion recognition through facial expression analysis based on a neurofuzzy network", Neural Networks, ELSAVIER, 18(4):423-35, May 2005.
[2] Alexander I. Iliev, Monograph "Emotion Recognition From Speech", Lambert Academic Publishing, ISBN-10: 3847377604, ISBN-13: 978-3847377603, 2012.
[3] Iliev, A.I., Scordilis, M.S., "Spoken Emotion Recognition Using Glottal Symmetry", EURASIP Journal on Advances in Signal Processing, Volume 2011, Article ID 624575, 2011.
[4] Igor Bisio, Alessandro Delfino, Fabio Lavagetto, Mario Marchese, and Andrea Sciarrone, "Gender-Driven Emotion Recognition Through Speech Signals for Ambient Intelligence Applications", IEEE Transactions On Emerging Topics In Computing, Vol. 1, No. 2, pp. 244-257, December 2013.
[5] Ververidis D. and Kotropoulos C., "Emotional Speech Recognition: Resources, Features, and Methods", Speech Communication, Vol. 48, pp. 1162-1181, 2006.
[6] Cowie R. and Cornelius R., "Describing the Emotional States that are Expressed in Speech", Speech Communication, Vol. 40, pp. 5-32, 2003.
[7] Cornelius R., "The Science of Emotion. Research and Tradition in the Psychology of Emotion", Upper Saddle River, NJ: Prentice-Hall, pp. 260, 1996.
[8] Bhatti M., Wang Y., and Guan L., "A Neural Network Approach for Human Emotion Recognition in Speech", IEEE International Symposium on Circuits and Systems, Vancouver BC, pp. 181-184, 2004.
[9] Noda T., Yano Y., Doki S., and Okuma S., "Adaptive Emotion Recognition in Speech by Feature Selection Based on KL-divergence", IEEE International Conference on Systems, Man, and Cybernetics in Taipei, Taiwan, pp. 1921-1926, October 8-11 2006.
[10] Kwon O., Chan K., Hao J., and Lee T., "Emotion Recognition by Speech Signals", Eurospeech Geneva, pp. 125-128, 2003.
[11] Go H., Kwak K., Lee D., and ChunM., "Emotion Recognition from the Facial Image and Speech Signal", SICE Annual Conference in Fukui, pp. 2890-2895, August 4-6 2003.
[12] Beckman M. and Elam G., "Guidelines for ToBI labeling", The Ohio State University Research Foundation, 1997.
[13] Gobl C. and Chasaide A., "The Role of Voice Quality in Communicating Emotion, Mood and Attitude", Speech Communication, Vol. 40, pp. 189-212, 2003.
[14] Iida A., Campbell N., Higuchi F., and Yasumara M., "A Corpus-Based Speech Synthesis System with Emotion", Speech Communication, Vol. 40, pp. 161-18, 2003.
[15] Iliev, A.I., Scordilis, M.S., Papa J.P., Falcão A.X., "Spoken emotion recognition through optimum-path forest classification using glottal features", Computer Speech and Language, ELSEVIER, Vol. 24, Issue 3, pp. 445-460, 2010.
[16] Daniel Neiberg, Kjell Elenius & Kornel Laskowski, "Emotion Recognition in Spontaneous Speech Using GMMs", INTERSPEECH 2006 – ICSLP, Pittsburgh Pennsylvania, pp. 809-812, September 17-21 2006.
[17] Stibbard, R., "Automated Extraction of ToBI Annotation Data from the Reading/Leeds Emotional Speech Corpus", ITRW on Speech and Emotion, ISCA in Speech Emotion, pp. 60-65, 2000.
[18] Evelyn R. M. Abberton, David M. Howard & Adrian J. Fourcin, "Laryngographic assessment of normal voice: A tutorial", Clinical Linguistics & Phonetics, Vol. 3, Issue 3, pp. 281-296, 1989.
[19] Witten, I., Frank, E., "Data Mining: Practical machine learning tools and techniques", Second edition, Morgan Kaufmann, 2005.
[20] Iliev, A.I., Zhang, Y., Scordilis, M.S., "Spoken Emotion Classification Using ToBI Features and GMM", Proceedings of the 14th International Workshop on Signals and Image Processing 2007 and the 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services, IEEE-IWSSIP 2007, Maribor Slovenia, pp. 495-498, June 27-30 2007.
[21] Markus, H., & Kitayama, S., "Culture and the self: Implications for cognition, emotion, and motivation", Psychological Review, 98(2), pp. 224–253, 1991.
[22] Perunovic, W., Heller, D., & Rafaeli, E., "Within-person changes in the structure of emotion: The role of cultural identification and language", Psychological Science, 18, pp. 607–613, 2007.
[23] Yuki, M., Maddux, W. W., & Masuda, T., "Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States", Journal of Experimental Social Psychology, 43(2), pp. 303–311, 2007.