Emotion Recognition: Differences between Spontaneous Dialogue and Active dialogue.

Divya Gupta, Poonam Bansal, Kavita Choudhary

Jagan Nath University, Jaipur India

Abstract: *In this paper, we study how aspects of a given database influence the performance of different approaches to emotion recognition, we performed cross experiments on the spontaneous dialogue database AVEC2012 and the dialogue database activated IEMOCAP. We have discovered that there is generally more DIS-NV in spontaneous dialogue than in acted dialogue. We also found complex variations in nonverbal vocalizations in spontaneous dialogue that were overlooked when designing data collection through play. Based on the distributions of the characteristics of Global Prosodic, the distribution of volume and the quality of the voice in the activated dialogue has wider ranges of values than in the spontaneous dialogue, and there is more variation in pitch in the activated dialogue than in the spontaneous dialogue. We found that the IEMOCAP database which noted emotions at the expression level benefited less from the use of the LSTM model than the AVEC2012 database which noted emotions at word level. This indicates that the ability of the LSTM to model a long-term time context may be more useful for emotion recognition tasks on a small time scale rather than on a time scale large.*

Keywords: **SVM, DIS-NV, Cross-Validation, spontaneous and acted dialogue.**

1. INTRODUCTION

Most of the earlier work on emotion recognition has focused on experiences using a single database)). Only a few studies have carried out cross-experiments to test the robustness of the characteristics or models proposed. Previous cross-studies on emotion recognition suggest that it is often difficult to generalize the effectiveness of features and models in different databases, especially when the type of dialogue is different. For example, Eyben et al. (2015b) carried out experiments on the 6 most used emotion databases, and their results showed that the performance ranking of 7 sets of standard acoustic characteristics varies considerably between the databases. Likewise, Schuller et al. (2010a) built an SVM model with LLD functionality to detect binary excitation and valence, and tested the recognizer of emotions in several databases, including spontaneous and active dialogue. The results of Schuller et al. (2010a) illustrate the great influence of the type of dialogue and the difficulty of predicting emotions in spontaneous dialogue. The investigation by Zeng et al. (2009) also highlighted the fact that the emotions in the acted dialogue are more acoustically exaggerated than the emotions in the spontaneous dialogue, which raises the question whether the performance of the emotion recognizers trained in the acted dialogue can decrease considerably when applied to a more natural setting. Therefore, they suggested collecting databases of more spontaneous dialogue emotions, which would lead to emotion recognizers that could be better generalized to natural interaction scenarios.

1.1 Distribution of Emotion Annotations

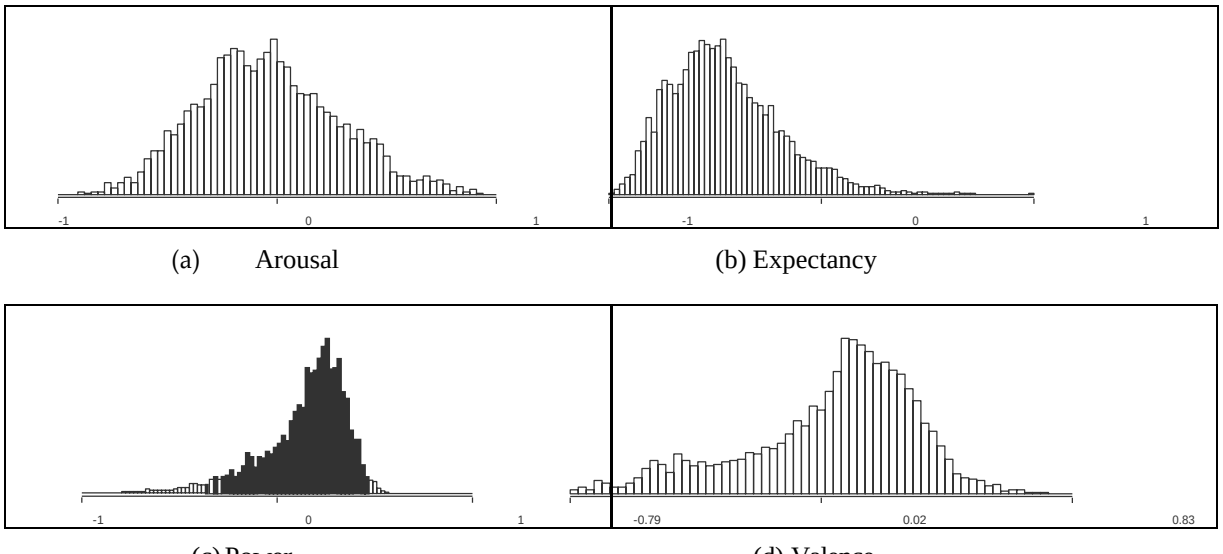
To study the difference between spontaneous and acted dialogue, we compared the distribution of emotional annotations in the spontaneous AVEC2012 database and the acted IEMOCAP database. The AVEC2012 and IEMOCAP databases noted emotions with different patterns. The AVEC2012 database scored emotions at the word level, while the IEMOCAP database scored emotions at the

expression level. We keep the annotation of emotions at the word level of the AVEC2012 database when we study the distribution of emotions and perform emotion recognition experiments. However, by studying the DIS-NV distributions and the acoustic characteristics, we reduced the sample of AVEC2012 data from the word level to the expression level and plotted descriptive statistics at the expression level for the two databases.

1.2 Distribution of Emotions in Spontaneous Dialogues

Figure 1 illustrates the distributions of the original continuous emotion annotations at the word level in the AVEC2012 database. Figure 2 illustrates the distributions of the discrete emotion annotations transformed at the word level in the AVEC2012 database. For each emotion dimension in Figure 2, the three bars from left to right respectively represent the low (dark blue), medium (red) and high (light blue) categories.

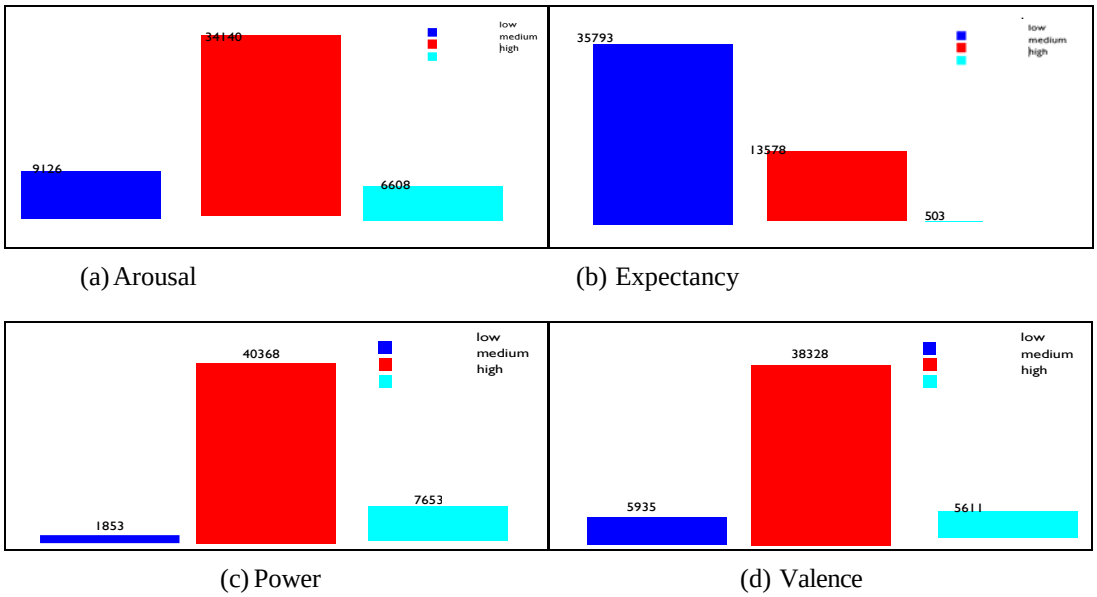
As we can see, for the emotional dimensions of excitement, power and valence, the middle category is significantly larger than the other two categories of emotion. This indicates that the emotions in spontaneous dialogue are soft or neutral most of the time, which is consistent with the previous finding that strong emotions are difficult to induce in spontaneous dialogue. Another interesting observation is that most of the data was rated as low or medium expectation, indicating that speakers often show signs of uncertainty during spontaneous dialogue. This is consistent with the previous finding that DIS-NVs are common in spontaneous, unscripted conversations. When compiling the AVEC2012 database, four virtual agents with different personality conceptions were used. Using Prudence, the silent and neutral virtual agent, can increase the number of neutral or non-emotional data instances in AVEC2012 database. The peak of the power and valence distribution located on the positive values axis in Figure 1 also indicates that it is difficult to conceive of a credible virtual agent that induces negative emotions in the participants.



(a) Arousal (b) Expectancy

(c) Power (d) Valence

Figure 1:Word-Level Continuous Emotion Distribution on the Spontaneous AVEC2012 Database



(a) Arousal (b) Expectancy

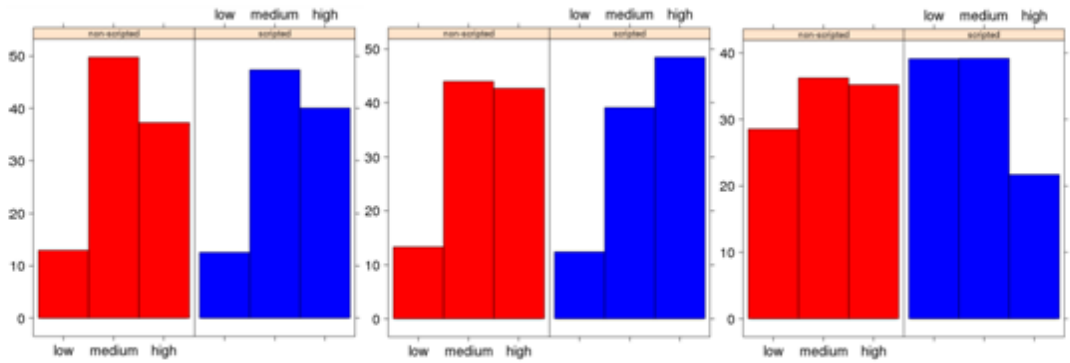
(c) Power (d) Valence

Figure 2: Word-Level Discrete Emotion Distribution on the Spontaneous AVEC2012 Database

1.4 Distribution of Emotions in Acted Dialogues

Figure 3 illustrates the distribution of emotion annotations at the level of the discrete expression transformed in the IEMOCAP database. In Figure 3, the red bars represent the instructions collected by acting without a hyphen and the blue bars represent the instructions collected by acting with a hyphen. The three columns from left to right on each graph respectively represent the low, medium and high categories of each emotional dimension. The y-axis in the figure is the percentage of the total number of data instances.

Compared to the distributions of emotions on the spontaneous AVEC2012 database (Figure 2), the categories of emotions are more balanced on the IEMOCAP database played. This indicates the advantage of collecting databases on emotions by acting, which means that the data is more balanced (Zeng et al., 2009). An interesting observation is that there are fewer statements with low excitation or low power than statements with medium and high excitation or power. This reflects the fact that when collecting the IEMOCAP database, the game scenarios were biased towards more active and dominant situations (for example, an intense argument to customer service). The emotion annotation distributions for utterances collected by scripted or unscripted actions are approximately the same, which is likely due to a similar scenario design in both cases. The difference between the annotation distributions of emotions on the AVEC2012 database and the IEMOCAP database indicates that spontaneous and acted dialogues are different in terms of the speaker's emotional status.



(a) Arousal (b) Power (c) Valence

1.6 Distribution of DIS-NVs

Figure 3: Utterance-Level Emotion Distribution on the Acted IEMOCAP Database

Here we study the differences between spontaneous dialogue and active dialogue in terms of DIS-NV in speech. As Trouvain (2014) suggests, DIS-NVs are more common in spontaneous and unscripted dialogue. Actually, the actors are trained to be fluent and the DIS-NVs are often not included in the scripts to collect the scripted dialogue. Therefore, we expect fewer expressions with DIS-NV in the IEMOCAP dialog-driven database than in the spontaneous AVEC2012 database.

To compare the occurrences of DIS-NV in the spontaneous and acted dialogue, we report the percentage of declarations containing each type of DIS-NV in the two databases in Table 1. "FP" represents a full break, "FL" represents padding, "ST" represents a stutter, "LA" represents a laugh, "AB" represents an audible breath. As we can see, reports with full pause, laughter and audible breathing are less frequent in the activated IEMOCAP database than in the spontaneous AVEC2012 database. This is in line with previous results. However, padding and stuttering are more common in unwritten expressions in the IEMOCAP database. This indicates that, compared to acting with a script, acting without a script is more like spontaneous dialogue in terms of the number of dysfluences in sentences.

Databases	FP(%)	FL(%)	ST(%)	LA(%)	AB(%)
AVEC2012	**32.0**	14.7	9.4	**11.9**	**2.7**
IEMOCAP (non-scripted)	14.9	**33.0**	**10.1**	2.4	0.8
IEMOCAP (scripted)	7.8	15.9	2.9	0.9	0.4

Table 1: Percentages of Utterances with DIS-NV in Spontaneous and Acted Dialogue

1.7 Additional DIS-NVs.

The DIS-NVs that we annotate are only a subset of all the DIS-NVs that occur in speech. Here we study the influence of the inclusion of other common DIS-NV types. Find out more in particular, we note speech repairs (SR, when the speaker is corrected), turn times (TT, silent pause at the start of a lap) and extensions (PL, prolonged pronunciation of a syllable) as DIS-NV additional to the IEMOCAP database. The percentage of reports containing these additional DIS-NVs is shown in Table 2. As we can see, compared to FP, FL and ST shown in Table 1, SR and PL are less frequent in the IEMOCAP database.

We also performed cross-validation experiments 10 times with an SVM model (C-SVC with RBF kernel) to compare the performances of our original DIS-NV set containing 5 DIS-NV (FP, FL, ST, LA, AB) and the extended DIS-NV package that includes the three additional DIS-NVs (FP, FL, ST, LA, AB, SR, TT, PL) in the IEMOCAP database. We report the results (F1 measurements) in Table 5.3. As we can see, adding these additional DIS-NVs does not improve the performance of emotion recognition. The original DIS-NV set of 5 DIS-NV constantly outperforms the extended DIS-NV set of 8 DIS-NV.1 Therefore, in later experiments, we continued to use the DIS-NV set of origin of 5 DIS-NV.

Databases	SR(%)	TT(%)	PL(%)
IEMOCAP (non-scripted)	**3.4**	27.9	**3.8**

IEMOCAP (scripted)	1.0	**35.1**	1.3

Table 2: Percentages of Utterance with Additional DIS-NV in IEMOCAP Database

Models	Arousal(%)	Power(%)	Valence(%)	Mean(%)
Original DIS-NV set	**36.3**	**40.7**	**32.8**	**36.6**
Expanded DIS-NV set	35.6	38.0	29.9	34.5

Table 3: Using Additional DIS-NVs for Emotion Recognition on IEMOCAP Database

1.8 Distribution of DIS-NVs in Spontaneous and Acted Dialogues

We have examined the complete pause and laughter distributions as examples of DIS-NV to explore the differences between spontaneous and acted dialogue. To study Regarding the distribution differences in more detail, we have plotted DIS-NV in the IEMOCAP database without hyphen and with hyphen separately. As an Expectation Emotion Dimension annotation is missing in the IEMOCAP database, we only trace here distributions full of pause and laughter in the dimensions Excitement, Power and Valence.

To compare the distribution in different types of dialog boxes, we draw smooth density graphs with lines representing the AVEC2012 database, the non-hyphenated subset of the IEMOCAP database and the subset with hyphen of the IEMOCAP database, respectively in Figures 4 and 5. In these figures, the x-axis represents the percentage of the total duration of an instruction that is a DIS-NV, the y-axis represents the percentage of instructions that have the value on the x-axis. We limit the maximum value of the y axis to 0.8 to zoom in on expressions containing DIS-NV. The histograms of all DIS-NV distributions in the two databases can be found in section C.1 of appendix C.

1.9 Distribution of Filled Pauses in Spontaneous and Acted Dialogues

As shown in Figure 4, the blue (IEMOCAP with hyphen) and green (IEMOCAP without hyphen) lines stop before the x axis reaches 0.68, while the pink line (AVEC2012) reaches 1.0 on the x axis. Keep in mind that in these figures, the x-axis represents the percentage of the total length of a paused sentence. This shows that there are no expressions containing more than 70% complete pause in the activated IEMOCAP database, while in the spontaneous AVEC2012 database, there are expressions which are complete pause. This reflects the fact that during the data collection, the IEMOCAP database used professional actors, who are trained to have fewer gaps than the general public involved in the data collection.

The filled breaks in the unscripted, hyphenated IEMOCAP dialog have similar distributions in Figure 4, except that there are fewer filled breaks in the hyphen performance dialog than in the dialog unscripted performance dialog or spontaneous dialogue (the blue line reaches to 0 in and axis before the green line in all the graphs in Figure 4). Although Busso et al. (2008) argue that acting without script is similar to spontaneous dialog, as we can see, there are fundamental differences in the full pause distributions between the dialog of spontaneous action (the red lines) and the unscripted (the green lines).

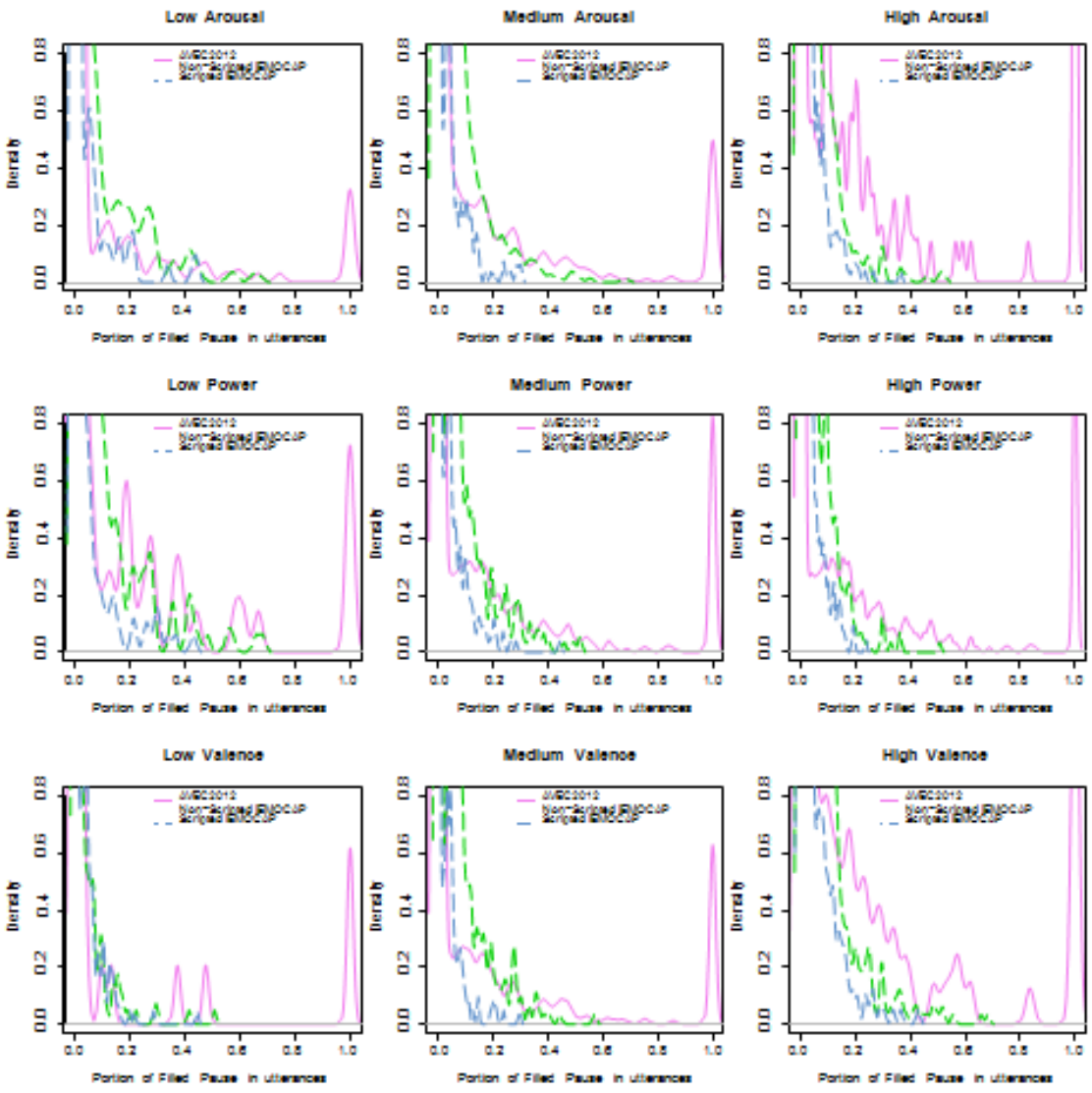


Figure 4: Filled Pause Distribution on Arousal, Power, Valence in Utterances

1.10 Distribution of Acoustic Features:

In addition to the distribution of speakers' and DIS-NV's emotions in speech, in this section, we study the acoustic differences between spontaneous and acted dialogue by studying the distribution of the global prosodic characteristics of Bone et al. (2014) in both databases. We study the IEMOCAP instructions separately with hyphen and without script for a more detailed understanding. Because the waiting dimension annotations are missing in the IEMOCAP database, we only compare the distributions in the dimensions of excitation, power and valence between the two databases in this section. The distributions of GP characteristics in the Hold dimension in the AVEC2012 database and the smoothed density graphs of all GP

characteristics in the two databases can be found in section C.2 of Appendix C.

1.10.1 Distribution of Log Pitch:

As shown in Table 4, the distribution of the central register tone is more skewed and has a lower standard deviation in spontaneous dialogue than in dialogue without hyphens or hyphens. These observations indicate that the distribution of mean register tone values has more variation in the acting dialogue than in the spontaneous dialogue.

Databases	Mean	Standard Deviation	Skewness
AVEC2012	0.412	**0.172**	**-0.360**
IEMOCAP (non-scripted)	0.487	0.195	-0.190
IEMOCAP (scripted)	0.480	0.199	-0.191

Table 4: Distribution of Log Pitch on AVEC2012 and IEMOCAP Databases

1.10.2 Distribution of Intensity

As shown in Table 5, the average intensity distribution is more skewed and has a lower standard deviation in spontaneous dialogue than in dialogue performed without hyphens or hyphens. This indicates that the volume of speech has a wider range in active dialogue than in spontaneous dialogue.

Databases	Mean	Standard Deviation	Skewness
AVEC2012	0.581	**0.095**	**-0.647**
IEMOCAP (non-scripted)	0.445	0.134	0.073
IEMOCAP (scripted)	0.455	0.134	0.232

Table 5: Distribution of Intensity on AVEC2012 and IEMOCAP Databases

1.10.3 Distribution of Voice Quality

Remember that the HF500 is calculated as the ratio of total energy greater than 500 Hz to low frequency energy in a statement. As shown in Table 5.6, similar to the intensity, the HF500 distribution is more asymmetrical and has a lower standard deviation in spontaneous dialogue than in dialogue performed without hyphens or hyphens. This indicates that there are more variations in voice quality in the acting dialogue than in the spontaneous dialogue.

Databases	Mean	Standard Deviation	Skewness
AVEC2012	0.570	**0.395**	**0.750**
IEMOCAP (non-scripted)	0.542	0.506	0.325
IEMOCAP (scripted)	0.400	0.463	0.747

Table 6: Distribution of Voice Quality on AVEC2012 and IEMOCAP Databases

2. Experiment 1: Influence of Dialogue Type on Effectiveness of DIS-NV Features

In the previous sections, we illustrated the differences in the DIS-NV distribution and the acoustic variations in spontaneous and acted dialogue. Our statistical analyzes show that compared to the spontaneous dialogue, in the activated dialogue, there is less DIS-NV and more acoustic variation. In this section, we have carried out emotion recognition experiments in the spontaneous databases AVEC2012 and IEMOCAP to study how these differences between spontaneous dialogues and acted dialogues influence the effectiveness of the emotion recognition functions in spoken dialogue.

2.1 Methodology

To study the influence of the type of dialogue on the effectiveness of the DIS-NV characteristics, we carried out the recognition of emotions both in the spontaneous database AVEC2012 and in the performed IEMOCAP database, and compared the experimental results (Tian et al., 2015a). We performed cross-validation experiments 10 times for classification experiments and reporting of weighted F measures to avoid the problem of unbalanced data. We assessed the importance of the differences in performance using the paired permutation test (Menke and Martinez, 2004) with 100,000 randomizations.

2.2 Results

The results of the models for the recognition of unimodal emotions from the AVEC2012 database are presented in Table 7. The results of the unimodal emotion recognition models in the IEMOCAP database are presented in Table 8. "Average" represents the arithmetic mean of the results in all emotional dimensions. Note that the IEMOCAP database did not provide wait annotations, so the results are missing in the wait dimension for the IEMOCAP database. We include a reference model that predicts the majority class.
We also observed that the acoustic characteristics are more predictive of emotions than the lexical characteristics in the actuated IEMOCAP database, while the lexical characteristics are more predictive than the acoustic characteristics in the spontaneous AVEC2012 database. Our results indicate that the effectiveness of the characteristics largely depends on the specific task of recognizing emotions, in particular the type of dialogue.

Models	Arousal(%)	Expectancy(%)	Power(%)	Valence(%)	Mean(%)
Baseline	51.6	55.6	66.4	58.8	58.1
AVEC-LLD	52.4	60.8	67.5	59.2	60.0
IS10-LLD	52.9	60.8	67.6	59.2	60.1
eGeMAPS	56.9	60.1	73.4	66.8	64.3
GP	56.3	60.0	72.4	66.8	63.9
DIS-NV	55.9	**61.4**	**74.7**	66.8	**64.7**
PMI	55.7	60.7	73.0	66.8	64.0
CSA	**57.5**	59.8	73.0	**67.1**	64.4

Table 7: Unimodal Emotion Recognition with SVM on the Spontaneous AVEC2012 Database

Models	Arousal(%)	Expectancy(%)	Power(%)	Valence(%)	Mean(%)
Baseline	31.7	#	28.7	27.0	29.1
LLD	**65.2**	#	**53.8**	**53.5**	**57.5**
eGeMAPS	60.9	#	52.2	49.4	54.1
GP	57.0	#	49.7	41.5	49.4
DIS-NV	36.3	#	40.7	32.8	*36.6*
PMI	47.8	#	48.1	32.9	42.9
CSA	47.0	#	47.2	29.5	41.2

Table 8: Unimodal Emotion Recognition with SVM on the Acted IEMOCAP Database

3. Experiment 2: Using Deep Learning for Unimodal Emotion Recognition

Here we studied the gain of using the deep and contextual LSTM model instead of the superficial and non-contextual SVM model. We have constructed unimodal models for recognizing LSTM emotions in the two databases.

3.1 Methodology

The performances of the LSTM unimodal models in the AVEC2012 and IEMOCAP databases are presented in tables 9 and 10. "Average" represents the arithmetic average of the results in the four dimensions of emotion. LSTM models with a single hidden layer are used when building these unimodal models and the number of memory cells was selected based on cross-validation experiments. We have included a basic model which predicts the class majority. We conducted cross-validation experiments 10 times in both databases and reported the F-weighted measures, and assessed the significance of the performance differences using the matched permutation test with 100,000 randomizations.

3.2 Results and Discussion

As shown in Table 9, based on our previous results, the DIS-NV characteristics predict emotions in spontaneous dialogue, in particular to predict the emotional dimension of expectation, 5 whether they are used with the SVM model or the LSTM model. The lexical characteristics of CSA benefit more from the use of the contextual LSTM model compared to the non-contextual SVM model and achieve the best overall performance in the AVEC2012 database.6 Compared to the performance of the SVM models presented in Table 7, the LSTM models improve performance in all emotional dimensions using each set of features. This indicates the effectiveness of the deep and contextual LSTM model for the recognition of emotions in spoken dialogue.

As shown in Tables 7 and 9, the IS10-LLD package produces similar or improved performance for the SVM7 and LSTM8 models compared to the AVEC-LLD package.

As shown in Table 10, according to our previous findings, the effectiveness of the characteristics varies when the type of dialogue is different. Compared to the performance of the SVM models presented in Table 8, the LSTM models improve performance in all emotional dimensions by using the smallest sets of GP, DIS-NV, PMI and CSA functions in the database. IEMOCAP. However, the performance of the LSTM

model with LLD and the eGeMAPS feature set is worse than the SVM models. This may be due to the fact that the LLD and eGeMAPS feature sets have higher dimensionality (1582 for LLD, 88 for eGeMAPS), resulting in more complex LSTM models that have more parameters to optimize during training. There are fewer training instances in the IEMOCAP database than in the AVEC2012 database (around 10,000 for IEMOCAP, around 50,000 for AVEC2012), which may limit the optimization of LSTM models that use the LLD and eGeMAPS feature set.

Another reason why the LSTM model does not offer significant performance gains may be that the IEMOCAP database recorded data at the instruction level, while the AVEC2012 database recorded data at the instruction level word. The long-term time context that the LSTM model tries to incorporate may not be useful or necessary when the time scale of data instances is as long as an instruction. We also observed that the characteristics inspired by knowledge work better than the statistical characteristics LLD in spontaneous and acted dialogue in most cases, which is consistent with previous studies on the recognition of emotions.

Models	Arousal(%)	Expectancy(%)	Power(%)	Valence(%)	Mean(%)
Baseline	51.6	55.6	66.4	58.8	58.1
AVEC-LLD	56.5	61.6	72.1	66.4	64.2
IS10-LLD	57.1	61.4	72.7	67.1	64.6
eGeMAPS	56.2	60.3	72.6	66.8	64.0
GP	56.0	60.3	72.4	66.8	63.9
DIS-NV	56.2	**65.9**	72.8	67.3	*65.5*
PMI	56.0	62.7	72.3	66.7	64.4
CSA	**58.1**	61.7	**75.2**	**70.2**	**66.3**

Table 9: Unimodal Emotion Recognition with LSTM on the Spontaneous AVEC2012 Database

Models	Arousal(%)	Expectancy(%)	Power(%)	Valence(%)	Mean(%)
Baseline	31.7	#	28.7	27.0	29.1
LLD	53.7	#	46.2	38.6	46.2
eGeMAPS	**60.1**	#	**52.2**	**46.6**	**53.0**
GP	58.0	#	50.6	41.8	50.1
DIS-NV	41.6	#	37.8	34.0	37.8
PMI	48.8	#	48.7	32.9	43.5
CSA	50.0	#	48.1	44.5	47.5

Table 10: Unimodal Emotion Recognition with LSTM on the Acted IEMOCAP Database

4. Discussion

In this paper, we discuss the differences between spontaneous dialogue and active dialogue. We have discovered that there is generally more DIS-NV in spontaneous dialogue than in acted dialogue. We also

found complex variations in nonverbal vocalizations in spontaneous dialogue that were overlooked when designing data collection through play. Based on the distributions of the characteristics of Global Prosodic, the distribution of volume and the quality of the voice in the activated dialogue has wider ranges of values than in the spontaneous dialogue, and there is more variation in pitch in the activated dialogue than in the spontaneous dialogue. The dialogue gathered by the unscripted performance shares similarities with spontaneous dialogue, while there are fundamental differences between scripted dialogue and spontaneous dialogue.

Our intersomatic experiences have shown that the DIS-NV characteristics are less predictive of emotions in the acted dialogue because there is less DIS-NV in the acted dialogue compared to the spontaneous dialogue. However, this chapter only considered models for the recognition of unimodal emotions.

We also studied the gain of using deep and contextual LSTM models compared to the use of superficial and non-contextual SVM models for the recognition of emotions. Our results show that although the LSTM model performs better than the SVM model in most cases, optimization of the complex LSTM model may be limited by the small amount of training data available. Therefore, it may be preferable to use the knowledge-inspired functionalities with the LSTM model which have a lower dimensionality and therefore fewer parameters to optimize. In the future, to further explore the benefits of using LSTM models for emotion recognition, we would also like to compare the performance of an LSTM model using built-in layers derived from data-based functionality low level with an LSTM model using inspiration inspired by Knowledge Direct Characteristics. In terms of gaining the inclusion of temporal contexts, we found that the IEMOCAP database which noted emotions at the expression level benefited less from the use of the LSTM model than the AVEC2012 database which noted emotions at word level. This indicates that the ability of the LSTM to model a long-term time context may be more useful for emotion recognition tasks on a small time scale (for example, frame or word level) rather than on a time scale large (for example, statement or level of conversation).

References:-

1. *Lindasalwa Muda, Mumtaj Begam and Elamvazuthi, I. (2010). Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. Journal Of Computing, Volume 2, Issue 3.*
2. *Sivaram, G.S.; and Hermansky, V.S. (2011). Multilayer perceptron with sparse hidden outputs for phoneme recognition. IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP), Prague, pp. 5336–5339.*
3. *Lindasalwa, Muda; Mumtaj, Begam; Elamvazuthi, I. (2010). Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic. Journal of Computer Theory and Engineering, Vol.2, No.6, 1793-8201.*
4. *Bansod, Nagsen , S; Siddharth, B; Dadhade, Seema S. Kawatheka.; and K. V. Kale. (2014). Speaker Recognition Using Marathi (Varhadi) Language. International Conference on Intelligent Computing Applications.*
5. *Rakesh, K.; Dutta S.; Shama K. (2011). Gender Recognition using speech processing techniques in LABVIEW. International Journal of Advances in Engineering & Technology,1(2): 51-63.*
6. *Ajay; Anadi; and Singh, P.K. (2015). Novel Digital Image Water Marking Technique Against Geometric Attacks, International Journal of Modern Education and Computer Science.*
7. *Furui S. (1983). Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-34.*
8. *Gaikwad, Santosh, Bharti Gawali, and Suresh Mehrotra. (2013). Creation of Marathi speech corpus for automatic speech recognition. International Conference Oriental COCOSDA held jointly with Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE).*
9. *Gaikwad, Santosh, Bharti Gawali, and S. C. Mehrotra. (2012). Novel approach based feature extraction for Marathi continuous speech recognition. Proceedings of the International Conference on Advances in Computing Communications and Informatics - ICACCI 12 ICACCI 12.*
10. *Awasthy; Saini, Neeta; Chauhan J.P.; D.S. (2005). Spectral analysis of speech: a new technique. International Journal of Signal Processing, Jan 2005 Issue.*
11. *Ooi Chia Ai, M. Hariharan,, Sazali Yaacob, Lim Sin Chee, 2012. Classification of speech dysfluencies with MFCC and LPCC features, Expert Systems with Applications, Vol.39 (2), 2157– 2165.*

12. Engin Avci , Zuhtu Hakan Akpolat, 2006. , Speech recognition using a wavelet packet adaptive network based fuzzy inference system, Expert Systems with Applications, Volume 31, Issue 3, pp. 495–503.

13. Vimal Krishnan V.R, Babu Anto P, 2009. Features of Wavelet Packet Decomposition and Discrete Wavelet Transform for Malayalam Speech Recognition, International Journal of Recent Trends in Engineering, Vol. 1(2), 93-96.

14. Yang Jie, 2009. Noise robust speech recognition by combining speech enhancement in the wavelet domain and Lin-log RASTA, ISECS International Colloquium on Computing, Communication, Control, and Management, IEEE Xplore, (Aug. 8-9, 2009), Vol. 2, 415-418.

15. Shivesh Ranjan, 2010. Exploring the Discrete Wavelet Transform as a Tool for Hindi Speech Recognition, International Journal of Computer Theory and Engineering, Vol. 2, No. 4, 642-646.

16. Sonia Sunny, David Peter S., K. Poulose Jacob. 2011, Wavelet Packet Decomposition and Artificial Neural Networks based Recognition of Spoken Digits, International journal of machine intelligence, Vol.3, issue 4, 318-321.

17. M.A.Anusuya, 2011. Comparison of Different Speech Feature Extraction Techniques with and without Wavelet Transform to Kannada Speech Recognition, International Journal of Computer Applications, Vol. 26, No.4, 19-24.

18. Picone J.W., 1993. Signal Modelling Technique in Speech Recognition, Proc. of the IEEE, Vol. 81, No.9, 1215-1247.

19. Rabiner L., Juang B. H., 1993. Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ.

20. Jeremy Bradbury, 2000. Linear Predictive Coding.

21. S. Mallat, 1999. A wavelet Tour of Signal Processing, Academic Press, San Diego.

22. K. P Soman, K.I Ramachandran, N.G Resmi,2010. Insight into Wavelets From Theory to Practice, PHI Learning Private Ltd, New Delhi.

23. Elif Derya Ubeyil, 2009. Combined Neural Network Model Employing Wavelet Coefficients for ECG Signals Classification, Digital signal Processing, Vol 19, 297-308.