DIS-NV Functions for the Recognition of Emotions in Spoken Dialogue

Divya Gupta, Poonam Bansal, Kavita Choudhary

Abstract*: We present our studies on the use of characteristics that describe the occurrences of DISfluence and nonverbal vocalization (DIS-NV) in spoken expressions for the recognition of emotions in 0"turn" to denote the continuous speech made by one speaker without interrupting the other speaker. Note that each speaker tower can contain one or more declarations, and consecutive speaker declarations may or may not belong to the same speaker tour. Here, our definition of speaker tower focuses on feeling and integrity in speech production, which differs from "tower" in the context of a tower system, which focuses on the transition between different speakers. We carried out experiments in the spontaneous dialogue database AVEC2012 to study the effectiveness of the proposed work. Our results show that our DIS-NV functions offer better performance than LLD or PMI functions in predicting all emotional dimensions. The DIS-NV characteristics are particularly predictive of the emotional dimension Waiting linked to the speaker's uncertainty and allow the best reported result to be obtained. The emotion recognition model using only the 5 DIS-NV functions achieved overall performance linked to the best reported result obtained by a multimodal emotion recognition model using thousands of audiovisual and lexical functionalities. These results confirmed that the proposed characteristics of DIS-NV are predictive of emotions in spontaneous dialogue.

***Keywords*: DIS-NV, spontaneous dialogue, DISfluence, cross-correlation score**

I. INTRODUCTION

In this study we explain DIS-NV and review the psycholinguistic studies on DIS-NV and emotions. We also explain our motivation to propose DIS-NV functions for the recognition of emotions in spoken dialogue.

A. Definition of DIS-NVs

Dissolutions are phenomena in discourse which "interrupt the flow of discourse and do not add propositional content to a statement" (Fox Tree, 1995). Previous psycholinguistic studies on dysfluences have focused on dysfluences caused by speech disorders. However, normal speech disorders have recently received increasing attention as they are common and important phenomena in dialogue, and have functions such as cornering retention (Lickley, 2015). Shriberg (2005) also argues that dysfluences are common in spontaneous speech and reflect the cognitive aspects of language production and interaction management. Psycholinguistic studies of spontaneous dialogue have shown that on average, for 100 words produced by the speaker, there are approximately 6 dysfluences (Finlayson, 2014). Speech production has three main phases: conceptualization, planning and articulation. Dissolutions can be generated in one of these three phases. For example, when the speaker organizes answers to a complex question (the conceptualization stage), when he searches for an appropriate word (the planning stage) or when he has trouble pronouncing a syllable (the stage articulation).).

Nonverbal vocalizations are sounds that the speaker produces in expressions other than verbal content. Dissolutions are sometimes included as types of non-verbal vocalization. However, it is more common to differentiate dysfluences from nonverbal vocalizations. There are two main types of nonverbal vocalizations: vocal qualifiers (for example, audible breathing or coughing) and vocal qualifiers (for example, laughing or crying). A cross-corpus study comparing six different corpora showed that among the different types of non-verbal vocalizations, laughter and audible breathing are the most frequent in spontaneous dialogue.

B. DIS-NVs and Emotions

The relationship between deficiencies and emotions has been neglected in previous psycholinguistic research. However, emotions can influence the neural mechanisms of the brain and therefore influence sensory processing and attention. This in turn influences speech processing and production, particularly the conceptual and planning stages of speech production, which can lead to deficiencies. Current studies on human dialogue also suggest that deficiencies convey information such as the level of conflict (Vidrascu and Devillers, 2005), the uncertainty of the speaker (Lickley, 2015) or the points of interest during meetings (Shriberg , 2005). Therefore, we expect more speech influence when the speaker is uncertain or when there is a hot spot in the dialogue. Although psycholinguistic studies suggest a possible relationship between DIS-NV and emotions, previous work on automatic emotion recognition has rarely used DIS-NV as input characteristics. To our knowledge, the only previous work using DIS-NV for the detection of emotions is the work of Vidrascu and Devillers (2005), which included the number of complete breaks per declaration ("uh" in French) in its set of functionalities Recognize 20 categories of emotions from the records of a French medical emergency call center. They compared the individual predictability of the characteristics and found that the complete break is the second most predictive characteristic (the range F0 of the statement is the most predictive characteristic).

Revised Manuscript Received on September 25, 2020.
***** Correspondence Author

Divya Gupta*, CSE Department, Jagan Nath University, Jaipur, India. Email: fromdivya81@gmail.com

Poonam Bansal, CSE department, GGSIPU University Delhi India. Email: pbansal89@gmail.com

Kavita Choudhary, Jagan Nath University, Jaipur India Email: kavita.yogen@gmail.com

However, they did not indicate the extent to which the emotional recognition model had benefited from the inclusion of the full pause function.

C. Types of DIS-NV

To extract our DIS-NV functions, we manually annotate three types of dysfluences, namely full breaks, padding and stuttering. We also use two types of nonverbal vocalizations provided in manual database transcripts, namely laughter and audible breathing. We focus on these specific types of DIS-NV as they are emotionally related and are most prevalent in spontaneous dialogue. We know that these five types of DIS-NV are only a subset of all the DIS-NV in speech. However, our experiences show that the inclusion of additional DIS-NVs does not improve emotional recognition performance. Therefore, in our emotion recognition experiments, the DIS-NV feature set contains the five selected DIS-NVs if they are not specified in another way.

• **Full Pauses:** non-lexical insertions in the speech used by the speaker when he stops to think while trying to keep his turn. For example, "Hmm" in the statement "Hmm. Maybe we should try another way." The three most common full pauses that we find in the AVEC2012 spontaneous dialogue database are "em", "eh" and "oh".

• Fillers: stuffed lexical breaks. For example, "you know" in the phrase "I just want to, you know, have a drink and forget everything." Some psycholinguistic studies do not distinguish between full breaks and full breaks (for example, Finlayson (2014)). In our work, we consider the breaks filled and filled separately to have a more detailed understanding of their relationship to emotions. The three most common charges we found in the AVEC2012 spontaneous dialog database are "good", "you know" and "I mean".

• Stuttering: words or part of a word that the speaker involuntarily repeats during the speech. For example, "Sa" in the statement "Sa. Saturday will be fine", or

First "I didn't do it" in the statement "I didn't do it, I didn't mean it".

• Laughter: a physical reaction that usually consists of rhythmic, often audible, contractions of the diaphragm and other parts of the respiratory system. Laugh annotations have been included in the manual transcripts provided with the two databases. Note that these are binary annotations of the presence / absence of laughter without differentiating the different types of laughter.

• Audible breathing: sounds generated by the movement of air through the respiratory system. Audible breath annotations have been included in the manual transcripts provided with the two databases.

D. Feature Extraction

We used a 15-word moving window to calculate the dysfluence characteristics for word-level emotion recognition in the AVEC2012 database. We chose a 15 word window because it is the average length of a sentence in the AVEC2012 database. In our later experiments on emotion recognition at the program level for the IEMOCAP database, we used the program duration instead of the moving window to calculate the DIS-NV functions. We also tried to use the length of the statement instead of the mobile word-level emotion recognition window in the AVEC2012 database. However, performance is worse than using the popup and is not included here.

As shown in Figure 1, the window includes the current word and the 14 history words that precede it, and slides from the start of a dialog session to the end. The characteristic value of the word w for DIS-NV type D (Dw) is calculated as the ratio between the total duration of DIS-NV type D which appears in the word window w (TD) and the total duration of the w word window including rests between words (Tw). This results in five DIS-NV functions for each word:

$$D_w = \frac{T_D}{T_w}$$

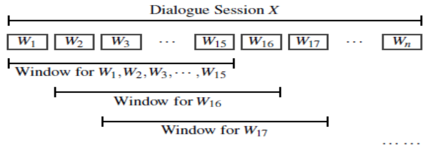
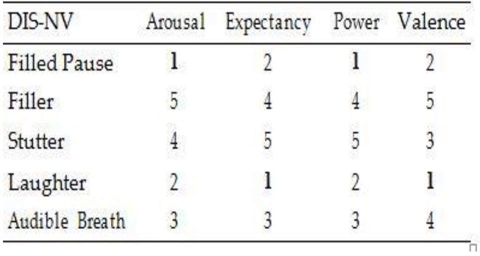


Figure 1: Window for Extracting DIS-NV Features from the AVEC2012 Database

E. INDIVIDUAL EFFECTIVENESS OF THE DIS-NV FEATURES

We compared the individual efficiency of DIS-NV functions in the AVEC2012 database using the correlation-based function selection method (CFS). The CFS method classifies the individual efficiency of functions depending on the recognition performance of each function It is used as a weak identifier, as well as the degree of redundancy between the characteristics. The results are presented in Table 1, with smaller numbers representing a higher classification by the CFS method and, therefore, greater individual efficiency. As we can see, pause and laughter are the most effective types of DIS-NV for emotions in spontaneous dialogue. The fact that the load is not highly classified supports the possibility of considering the complete break and the load separately to study the relationship between DIS-NV and emotions.



DIS-NV	Arousal	Expectancy	Power	Valence
Filled Pause	1	2	1	2
Filler	5	4	4	5
Stutter	4	5	5	3
Laughter	2	1	2	1
Audible Breath	3	3	3	4

Table 1: Individual Effectiveness Rankings of DIS-NV Features

F. Recognizing Emotions in Spontaneous Dialogue with DIS-NV Features

This study contains our experiences on the effectiveness of the DIS-NV functionalities proposed to recognize emotions in spoken dialogue. These include four experiences of emotional regression in the AVEC2012 spontaneous dialogue database. In experiment 1, we compared the performances of the DIS-NV functions with the reference acoustic and lexical characteristics. In Experiment 2, we studied the gain of incorporating DIS-NV characteristics with acoustic and lexical characteristics. In Experiment 3, we studied the influence of the temporal context for the recognition of emotions. In Experiment 4, we studied the automatic recognition of DIS-NV and the effectiveness of the self-detected DIS-NV functionalities for the recognition of emotions. Here we follow the AVEC2012 challenge protocol and conduct our experiments on the AVEC2012 spontaneous dialogue database.

II. EXPERIMENT 1: EMOTION REGRESSION IN SPONTANEOUS DIALOGUE WITH DIS-NV FEATURES

Our objective for Experiment 1 is to study the effectiveness of using DIS-NV functions by ourself for the recognition of emotions in spoken dialogue. To do this, we compared the performance of the proposed DIS-NV characteristics with reference acoustic and lexical characteristics and advanced emotion recognition studies.

A. Methodology

Emotions were noted at the word level as vectors of real value in the emotional dimensions of excitement, expectation, power and valence in the AVEC2012 database. . After setting up the AVEC2012 challenge, we use the cross-correlation score (CSC) as an evaluation measure. In the challenge, the AVEC2012 database is divided into three partitions, each with 32 dialogue sessions: the training partition, the development partition and the test partition. CCS is calculated as the average CC correlation coefficients between emotion predictions and emotion annotations during the 32 dialogue sessions of the test score of the AVEC2012 database:

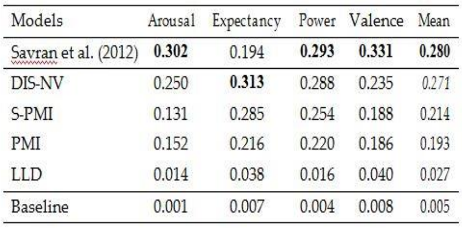
We evaluated the importance of the performance differences using the bilateral z test after Fisher's r to z transformation. The performances of the DIS-NV functionalities are compared to the acoustic LLD functionalities of reference AVEC2012 and to the lexical functionalities PMI. The acoustic characteristics of the LLD were provided in the AVEC2012 challenge as a set of reference characteristics, while the lexical characteristics of the PMI proved to be the most effective unimodal characteristic defined in previous work on this task (Savran et al. , 2012). We also compared the performance of our DIS-NV functions with the multimodal recognition results reported by other participants in the AVEC2012 challenge.

B. Results

The results of experiment 1 are reported in Table 2. "Average" represents the unweighted average of the results in the four dimensions of emotion. "DIS-NV" represents the model which uses the 5 proposed functions of DIS-NV. "S-PMI" represents the model using 1000 dispersed lexical characteristics of PMI, which was the most efficient set of functionalities in the previous works of the AVEC2012 database . “PMI” represents the model which uses the 8 non-dispersed PMI lexical characteristics that we have proposed. "LLD" represents the model which uses the basic LLD functionality 1842 AVEC2012. We use the support vector regression model to build all the emotion recognition models. We also include a reference model that predicts random numbers between [-1,1]. Since the evaluation metric is based on correlation, we cannot use a reference model that always predicts the mean.

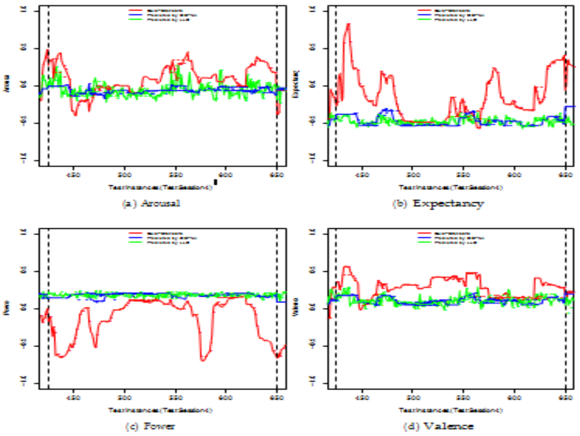
As shown in Table 2, our five knowledge-inspired DIS-NV characteristics obtained scores significantly higher than the reference acoustic and lexical characteristics for predicting each emotional dimension. Overall performance (average performance of all emotional dimensions) of the model that uses only our DIS-NV characteristics is linked to the best multimodal result reported in the AVEC2012 database. This indicates the effectiveness of the DIS-NV characteristics in recognizing emotions in spontaneous dialogue. The DIS-NV functionalities also obtained the best performances reported in the expected emotion dimension. This is consistent with the psycholinguistic conclusion that dysfluence is an indicator of speaker uncertainty (Lickley, 2015). Savran et al. (2012) surpassed the DIS-NV characteristics in the dimension of Valence, which may be due to the fact that Savran's model (2012) incorporated visual characteristics which describe facial expressions specifically effective in removing ambiguity from the dimension emotional of Valencia. The non-dispersed PMI characteristics that we propose have results close to the dispersed PMI characteristics of Savran et al. (2012) 3 while reducing the dimensionality of the functionalities from 1000 to 8. The LLD functionalities work extremely poorly here compared to the DIS-NV and PMI functionalities. This may be due to the high dimensionality and the frame-level nature of the LLD functionality compared to the DIS-NV and PMI functionality inspired by knowledge of the level of expression.

Table 2: Emotion Regression with DIS-NV Features on Spontaneous Dialogue



Models	Arousal	Expectancy	Power	Valence	Mean
Savran et al. (2012)	**0.302**	0.194	**0.293**	**0.331**	**0.280**
DIS-NV	0.250	**0.313**	0.288	0.235	*0.271*
S-PMI	0.131	0.285	0.254	0.188	0.214
PMI	0.152	0.216	0.220	0.186	0.193
LLD	0.014	0.038	0.016	0.040	0.027
Baseline	0.001	0.007	0.004	0.008	0.005

To further study the performance of the proposed DIS-NV functionalities, in Figure 2, we have drawn the predictions given by the DIS-NV and LLD functionalities in relation to the standard emotional annotations in Test Dialog 4 of the AVEC2012 database. As we can see, the predictions given by the characteristics of the LLDs are stronger and They have a flatter overall shape than DIS-NV functionality. For the predictions given by the characteristics of DIS-NV, there are segments which are straight lines due to the absence of DIS-NV in the declarations. However, when DIS-NVs are produced in the dialog, the general form of DIS-NV predictions better captures the form of the gold standard emotion annotations, the predictions having smaller absolute values than the gold standard annotations. The distributions of the different emotional dimensions vary considerably, indicating that the performance of models of emotion recognition in different emotional dimensions should be assessed separately and the mean SCC across all emotional dimensions should be considered as an additional benchmark. Note that the emotion annotations of the gold standard have a smoother shape than the automatic predictions in Figure 2. Therefore, it may be advantageous to use a sliding window to soften the predictions of the automatic emotion recognition model in the future.



(a) Arousal (b) Expectancy

(c) Power (d) Valence

Figure 2: Predictions vs. Annotations on the AVEC2012 Database

III. EXPERIMENT 2: MULTIMODAL EMOTION REGRESSION ON SPONTANEOUS DIALOGUE WITH DIS-NV FEATURES

In relation to the acoustic characteristics and the lexical content, the DIS-NVs contain additional information which can be linked to emotions. Therefore, in Experiment 2, we studied whether or not the emotion recognition model could benefit from the integration of our DIS-NV characteristics with reference acoustic and lexical characteristics.

A. Methodology

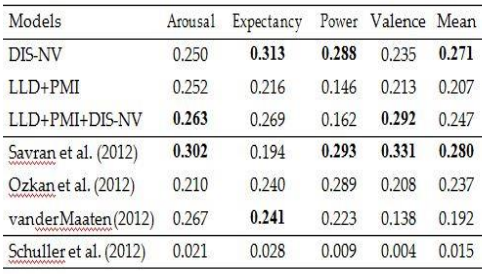
To study the gain of inclusion of our DIS-NV functions, we have built models for the recognition of multimodal emotions by concatenating sets of functions (i.e. fusion at the level of functions (FL)). As in experiment 1, we create support vector regression models for our unimodal and multimodal models and report CCS. The results of our experience are presented in Table 3 . "LLD + PMI" represents the model that uses the concatenated LLD feature set and our non-sparse PMI features.

B. Results

As Table 3 shows, the LLD + PMI + DIS-NV model performed much better than the LLD + PMI model in all emotional dimensions. The overall performance of the LLD + PMI + DIS-NV model is between the best (Savran et al. (2012)) and the second best model (Ozkan et al. (2012)) in the AVEC2012 challenge. The LLD + PMI + DIS-NV model also obtained the best result by predicting the emotional dimension waiting in relation to the reported challenge results. This verifies our conjecture that the DIS-NVs contain additional information on the acoustic characteristics or on the lexical content of the speech which is predictive of the emotions.

Therefore, the inclusion of DIS-NV functions in existing models improves performance. However, the LLD + PMI + DIS-NV model is less efficient than the DIS-NV unimodal model in the thrill of waiting and power. The reason may be the unbalanced size of the feature sets. In single entity concatenation (FL fusion), equal weights are assigned to each set of entities. Therefore, the highly predictive DIS-NV feature set with only 5 features can be overwhelmed by the noisy LLD feature set with more than a thousand features. This indicates that a better fusion strategy is needed to further improve the performance of the multimodal emotion recognition model.

Table 3: Multimodal Emotion Regression with DIS-NV Features on Spontaneous Dialogue



Models	Arousal	Expectancy	Power	Valence	Mean
DIS-NV	0.250	**0.313**	**0.288**	0.235	**0.271**
LLD+PMI	0.252	0.216	0.146	0.213	0.207
LLD+PMI+DIS-NV	**0.263**	0.269	0.162	**0.292**	0.247
Savran et al. (2012)	**0.302**	0.194	**0.293**	**0.331**	**0.280**
Ozkan et al. (2012)	0.210	0.240	0.289	0.208	0.237
vanderMaaten (2012)	0.267	**0.241**	0.223	0.138	0.192
Schuller et al. (2012)	0.021	0.028	0.009	0.004	0.015

C. Summary

Experiment 2 shows that DIS-NVs contain predictive information on emotions beyond the acoustic characteristics and the lexical content of spontaneous dialogue. However, the simple concatenation of functionalities can limit the gain of incorporating DIS-NV functionalities into multimodal emotion recognition models.

IV. EXPERIMENT 3: INFLUENCE OF CONTEXTUAL INFORMATION ON EMOTION REGRESSION

In Experiments 1 and 2, the DIS-NV functionalities and the PMI functionalities are expression level functionalities which include a time context, while the LLD functionalities are non-contextual frame level functionalities. Therefore, in Experiment 3, we study whether models based on LLD characteristics will benefit from the inclusion of the temporal context or not.

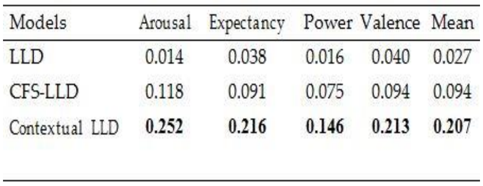
A. Methodology

To extract contextual LLD functionality, we first select a subset of the LLD functionality using the correlation-based subset of functionality selection, which gives 116 functionalities (the CFS-LLD functionality set). Then, we use the sliding window illustrated in figure 1 to calculate the minimum, maximum, average and standard deviations of the values of the CFS-LLD characteristics in the window, which gives 464 (116 × 4) LLD characteristics. contextual.

B. Results

The CCS of the unimodal models using the LLD characteristic set without original context, the LLD characteristic set without CFS context and the contextual LLD characteristic set are presented in Table 4. As we can see, the contextual characteristics of the LLD work much better than the non-contextual characteristics of the LLD and CFS-LLD to predict all the emotional dimensions. This verifies that the recognition of emotions can benefit by including the temporal context, which is consistent with the psychological results (Ortony et al., 1990). The improvement achieved by the engineering of the CFS functionalities compared to the direct use of the LLD functionalities also indicates that learning a more abstract representation of the functionalities and the reduction in the dimensionality of the functionalities are useful for the recognition of emotions.

Table 4: Influence of Temporal Context for Emotion Regression on Spontaneous Dialogue



Models	Arousal	Expectancy	Power	Valence	Mean
LLD	0.014	0.038	0.016	0.040	0.027
CFS-LLD	0.118	0.091	0.075	0.094	0.094
Contextual LLD	**0.252**	**0.216**	**0.146**	**0.213**	**0.207**

V. EXPERIMENT 4: AUTOMATIC DETECTION OF DIS-NVS

In this study, we focus on the DIS-NV functionalities based on manual DIS-NV annotations (Gold Standard DIS-NV functionalities) because we are interested in the effectiveness of DIS-NV for the recognition of emotions in spoken dialogue. . However, beyond improving the state of the art of emotion recognition in spoken dialogue, our long-term goal is to improve the quality of emotional interaction in HCI systems. In a fully automatic emotion recognition model, DIS-NV functionality will need to be extracted automatically, which can introduce noise into the DIS-NV functionality set. Therefore, in Experiment 4, we conducted a preliminary study on the influence of the use of the self-detected DIS-NV characteristics for the recognition of emotions. Note that automatic detection of speech and nonverbal vocalizations is an active area of research in itself. Therefore, with the improved DIS-NV recognition models, we will be able to further reduce the difference between self-detected DIS-NV functionality and Gold standards in the future.

A. Review on Automatic Detection of DIS-NVs

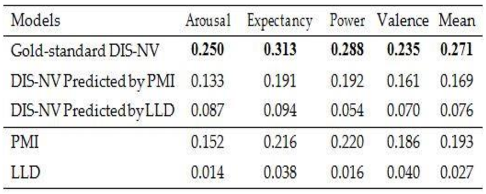
The automatic detection of DIS-NV has aroused the interest of voice recognition and psycholinguistic communities. DIS-NV detection models can improve the performance of automatic speech recognition and help researchers understand the process of voice generation.

For the automatic detection of dysfluence, various acoustic functionalities and machine learning algorithms were applied. Among the different acoustic characteristics, the tone and the duration were identified as highly predictive of dysfluences. For example, O'Shaughnessy and Gabrea (2000) classified complete pauses as vowels with durations greater than 120 ms and F0 less than the speaker's average F0. The stability of formants is also used by Audhkhasi et al. (2009) and Barczewska and Igras (2013) to detect dysfluences. In addition to tone, cepstral characteristics, such as the cepstral frequency coefficients Mel (MFCC), are also widely used in previous work (for example, Stouten and Martens (2003)). Previous studies on the detection of dysfluence have shown that contextual models are often powerful for the detection of dysfluence. For example, Yu et al. (2012) combined a hidden Markov model with a deep neural network and reached an error rate of 16.1% to detect dysfluence. Likewise, Zayats et al. (2016) built a bidirectional LSTM model and achieved new generation dysfluence detection performance with an F1 measurement of 85.9%. For the automatic detection of nonverbal vocalizations, most of the above Research has focused on the detection of binary laughter. Previous work on automatic laughter detection has studied various types of acoustic characteristics, such as prosodic characteristics (Truong and Van Leeuwen, 2007) and MFCC (Krikke and Truong, 2013). Like the detection of dysfluence, among the different types of acoustic characteristics, the tone was found to be very predictive of laughter (Salamin et al., 2013). Parallenguistic studies have shown that F0 in laughter is greater than F0 in speech segments (Rothgänger et al., 1998; Bachorowski et al., 2001). For the detection of audible breath, previous work has focused on prosodic (e.g. Braunschweiler and Chen (2013)) and cepstral (e.g. Ruinskiy and Lavner (2007)) features. Various machine learning algorithms have been applied to detect nonverbal vocalizations, such as the Gaussian mixing models used by Krikke and Truong (2013), the multilayer perceptrons used by Knox and Mirghafori (2007) and the vector-supported machines used . by Kennedy and Ellis (2004) Dupont et al. (2016) achieved peak laughter detection performance with 79% accuracy by combining audio and visual information.

B. Effectiveness of Auto-Detected DIS-NV Features for Emotion Recognition

Research is underway on automatic detection of dysfluence (for example, Liu et al. (2006)) and nonverbal vocalizations (for example, Niewiadomski et al. (2013)) . In this study, we focus on the performance of the Gold Standard DIS-NV functionality for emotion recognition. Here, we are conducting a preliminary experiment on the influence of automatic detection of DIS-NV. We use undispersed PMI and reference LLD characteristics AVEC2012 with an SVM model to predict the values of the DIS-NV characteristic. The CCS of emotion recognition models using the self-detected DIS-NV functions is presented in Table 5.

Table 5: Using Auto-Detected DIS-NV Features for Emotion Regression on Spontaneous Dialogue



Models	Arousal	Expectancy	Power	Valence	Mean
Gold-standard DIS-NV	**0.250**	**0.313**	**0.288**	**0.235**	**0.271**
DIS-NV Predicted by PMI	0.133	0.191	0.192	0.161	0.169
DIS-NV Predicted by LLD	0.087	0.094	0.054	0.070	0.076
PMI	0.152	0.216	0.220	0.186	0.193
LLD	0.014	0.038	0.016	0.040	0.027

As shown in Table 5, the performance of the self-detected DIS-NV features has a significant decrease in all dimensions of emotion compared to the standard DIS-NV features. However, the two self-detected DIS-NV functions still work much better than the reference LLD functions AVEC2012 for the recognition of emotions in spontaneous dialogue. Our results suggest that in addition to acoustic characteristics, lexical characteristics are also powerful predictors of DIS-NV. Note that we used a naive DIS-NV recognizer in this experiment. With an improved DIS-NV detection model, the performance difference between self-detected DIS-NV features and standard DIS-NV features can be further reduced. For example, Shi (2016) studied the auto-detection of dysfluences in the AVEC2012 database, Wang (2016) studied the auto-detection of nonverbal vocalizations in the AVEC2012 database.Shi (2016) used an automatic encoder on the characteristics of eGeMAPS and built an LSTM model with representation of the coded characteristics for the detection of dysfluence. The highest F1 measurements achieved are total rupture = 77.0%, filling = 78.0%, stuttering = 80.0%.

Wang (2016) built a network of deep beliefs with restricted layers of Boltzmann machines from Berninary Binary and combined the eGeMAPS function set with 78 MFCC functions for the detection of nonverbal vocalizations.

The highest F1 measurements achieved are: laughter = 69.8%, audible breathing = 78.6%. These results indicate that DIS-NV automatic detection can be performed with stable performance. Therefore, the self-detected DIS-NV characteristics will continue to be predictive of emotions in spontaneous dialogue.

C. Summary

In experiment 4, we showed that DIS-NV in spontaneous dialogue can be detected automatically with stable precision, and the characteristics of self-detected DIS-NV are always predictive of emotions in spontaneous dialogue. This indicates that our emotion recognition model using DIS-NV functions has the potential to apply to a fully automatic HCI system in the future.

VI. DISCUSSION

In this study, we have proposed DIS-NV functions for recognizing emotions. We explain the use of DIS-NV for the recognition of emotions and describe the calculation of DIS-NV features. We carried out experiments in the spontaneous dialogue database AVEC2012 to study the effectiveness of the proposed DIS-NV characteristics compared to the acoustic and lexical reference characteristics widely used in previous works of the same database.

Our results show that our DIS-NV functions offer better performance than LLD or PMI functions in predicting all emotional dimensions. The DIS-NV characteristics are particularly predictive of the emotional dimension Waiting linked to the speaker's uncertainty and allow the best reported result to be obtained. The emotion recognition model using only the 5 DIS-NV functions achieved overall performance linked to the best reported result obtained by a multimodal emotion recognition model using thousands of audiovisual and lexical functionalities. These results confirmed that the proposed characteristics of DIS-NV are predictive of emotions in spontaneous dialogue.

Our experience with incorporating DIS-NV functionality with other acoustic and lexical functionality indicates that DIS-NV contains additional emotional information in relation to acoustic functionality and lexical content. However, a better merge strategy than a simple concatenation of functionality is necessary to increase the gain of the merge mode. We also verified that including the temporal context is beneficial for the recognition of emotions. In addition, we have conducted preliminary experiments which have demonstrated that DIS-NV can be detected automatically with robust accuracy, so that DIS-NV functionality can continue to be effective in a fully automatic emotion recognition model.

One thing to note is that the evaluation metric based on the correlation coefficient reported by all of our models and all previous work using continuous emotion annotations WITH2012 is extremely weak. This indicates that emotion recognition is a task difficult. We assigned the original continuous emotion annotation from the AVEC2012 database into three discrete categories for each emotion dimension: low (original values in the range [-1, - 0.333)), medium (original values in the range [-0.333, + 0.333]) and high (original values in the range (+ 0.333, + 1]).

REFERENCES

1. Akansha Madan, Divya Gupta, “.Speech Feature Extraction and Classification: A Comparative Review”, International Journal of computer applications, (0975-8887) Volume 90 – No 9, March 2014 2014.
2. Y. Yuan, P. Zhao, and Q. Zhou, "Research of speaker recognition basedon combination of LPCC and MFCC," in Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on,2010, pp. 765-767
3. Kevin M. Indrebo, Richard J. Povinelli, Michael T. Johnson, IEEE Minimum Mean-Squared Error Estimation of Mel-Frequency Cepstral Coefficients Us-ing a Novel Distortion Model IEEE Transactions On Audio, Speech, And Language Processing, Vol. 16,No. 8, November 2008.
4. IEEE Trans. Audio Speech Lang. Process., vol. 16, no. 1, pp. 65–73, Jan. 2008.
5. L. S. Chee, Ooi Chia Ai, and S. Yaacob, "Overview of Automatic Stuttering Recognition System," in International Conference on Man-Machine Systems (ICoMMS 2009) Penang, Malaysia, 2009.
6. 6.K. Hu and D. Wang, ‘‘Unvoiced speech segregation from nonspeech inter-ference via CASA and spectral subtraction,’’ IEEE Trans. Audio Speech Lang. Process., vol. 19, no. 6, pp. 1600–1609, Aug. 2011.
7. Luengo and E. Navas, “Feature analysis and evaluation for automatic emo-tion identification in speech”, IEEE Trans.on Multimedia, vol. 12,no. 6, pp. 267-270, Oct 2010.
8. M. G. Sumithra and A. K. Devika, "A study on feature extractiontechniques for text independent speaker identification," in Computer Communication and Informatics (ICCCI), 2012 International Conference on, 2012, pp. 1-5.
9. Biswas, S. Ahmad, and M. K. Islam Mollat, "Speaker Identification Using Cepstral Based Features and Discrete Hidden Markov Model," information and Communication Technology, 2007. ICICT '07.International Conference on, 2007, pp. 303-306.
10. Y. Yuan, P. Zhao, and Q. Zhou, "Research of speaker recognition basedon combination of LPCC and MFCC," in Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on,2010, pp. 765-767.