Emotion Recognition in Spontaneous and Acted Dialogues

Leimin Tian
School of Informatics
the University of Edinburgh
Edinburgh, UK, EH8 9AB
Email: s1219694@sms.ed.ac.uk

Johanna D. Moore
School of Informatics
the University of Edinburgh
Edinburgh, UK, EH8 9AB
Email: j.moore@ed.ac.uk

Catherine Lai
School of Informatics
the University of Edinburgh
Edinburgh, UK, EH8 9AB
Email: clai@inf.ed.ac.uk

Abstract—**In this work, we compare emotion recognition on two types of speech: spontaneous and acted dialogues. Experiments were conducted on the AVEC2012 database of spontaneous dialogues and the IEMOCAP database of acted dialogues. We studied the performance of two types of acoustic features for emotion recognition: knowledge-inspired disfluency and nonverbal vocalisation (DIS-NV) features, and statistical Low-Level Descriptor (LLD) based features. Both Support Vector Machines (SVM) and Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) were built using each feature set on each emotional database. Our work aims to identify aspects of the data that constrain the effectiveness of models and features. Our results show that the performance of different types of features and models is influenced by the type of dialogue and the amount of training data. Because DIS-NVs are less frequent in acted dialogues than in spontaneous dialogues, the DIS-NV features perform better than the LLD features when recognizing emotions in spontaneous dialogues, but not in acted dialogues. The LSTM-RNN model gives better performance than the SVM model when there is enough training data, but the complex structure of a LSTM-RNN model may limit its performance when there is less training data available, and may also risk over-fitting. Additionally, we find that long distance contexts may be more useful when performing emotion recognition at the word level than at the utterance level.**

Keywords*—*emotion recognition, disfluency, LSTM, dialogue

I. INTRODUCTION

Research in cognitive science has shown that emotions are vital in human cognition and communication processes [1], such as memory [2], decision making [3], and social behaviour [4]. Therefore, it is also important for research in Artificial Intelligence to model emotional intelligence. This led to the establishment of the field of Affective Computing, in which emotion recognition has been a focus. It has become increasingly apparent that automatic recognition of emotion is crucial for advancing technologies related to human-computer interaction, such as human-agent dialogue systems.

In a virtual agent dialogue system, the ability to recognize and express emotions can make the agent appear more natural and believable to its human dialogue partner. It also increases user satisfaction and task success rate. For example, a virtual agent that is able to copy and adapt its laughter and expressive behaviours has been shown to increase users' humour experience [5]. Similarly, in affective game design, Non-Player Characters that are aware of the emotional states of the player and can generate emotional reactions have been shown to keep players engaged and to improve their gaming experience [6]. In a teaching scenario, a robot lecturer expressing a positive mood while giving lectures increased the arousal and positivity of the audience, as well as its perceived lecturing quality [7].

State-of-the-art approaches for improving emotion recognition performance focus on identifying better feature representations and applying models that fuse multiple modalities and include contextual information. Previous studies proposed various features and models. However, the effectiveness of these approaches may vary for different emotion recognition tasks. How to choose from these approaches is still an open problem. In this work, we compare knowledge-inspired features which describe the occurrence of disfluencies and nonverbal vocalisations (DIS-NV) in utterances, with statistical features which describe acoustic characteristics of the data. We also compare the performance of the widely used Support Vector Machines (SVM) and the Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) as classification models. We conducted emotion recognition tasks on both spontaneous and acted dialogues to gain a better understanding of the impact of different types of data.

A. Background

Disfluency and non-verbal vocalisations are important phenomena in natural speech. To the best of our knowledge, there have been no psycholinguistic studies showing a direct relation between disfluencies and emotions. However, emotions can influence the neural mechanisms in the brain, and thus influence sensory processing and attention [8]. This in turn influences speech processing and production, which may result in disfluencies. Current studies on human-human dialogues suggest that disfluency conveys information such as level of conflict [9], or uncertainty of the speaker [10]. Non-verbal vocalisations, especially laughter, have also been identified as universal and basic cues in human emotion recognition [11].

In our previous work [12], we showed that DIS-NV based features obtained state-of-the-art performance for recognizing emotions in spontaneous dialogues, and were found to be the most predictive type of feature. Consistent with psycholinguistic studies, the DIS-NV features were especially predictive for the Expectancy dimension which relates to the uncertainty of the speaker. However, our recent experiments on the IEMOCAP database of acted dialogue indicate that the DIS-NV features may be less predictive in acted dialogues

[13]. Thus, we would like to investigate aspects of the data that influence the predictiveness of different features and models.

Previous work applies the classic regression or classification algorithms to build emotion recognition models, such as Support Vector Machines [14], Hidden Markov Models [15], and Conditional Random Fields [16]. There have also been studies on feature engineering for emotion recognition, such as Canonical Correlation Analysis [17], and Correlation-based Feature-subset Selection [12]. There are many different algorithms to choose from and their effectiveness varies for specific tasks. However, previous work suggests that the predictiveness of features may have greater influence. That is, there may not be significant differences between the performance of different machine learning algorithms when using the same feature set under similar circumstances [18].

In recent years, deep learning models have obtained leading performance in machine learning tasks, especially in the areas of computer vision and speech recognition [19]. The network structure of deep learning models allows flexible control when fusing multiple modalities and including contextual information, which enables the models to learn better feature representations automatically. They have also achieved improved performance in emotion recognition compared to conventional machine learning algorithms. For example, deep hierarchical neural networks obtained the best reported results in detecting the Valence emotional dimension values and level of conflict [20], and the use of autoencoders has improved unsupervised domain adaptation in affective speech analysis [21].

However, compared to databases used for speech or image recognition tasks, the emotional databases are relatively small. This may limit optimization of the complex model structure of a deep learning model. The ability to generalize over different databases is also an issue for current deep learning models. In this work, we use the LSTM-RNN model as an example to investigate the predictiveness and robustness of deep learning models for emotion recognition, and compare their performance with the widely used SVM model.

B. Our Work

There are three important aspects for building an emotion recognition model: the data, the feature set, and the classification or regression model. For the data aspect, there are two main approaches for collecting conversational emotional databases: by recording acted or spontaneous dialogues.

For the feature aspect, there are two main types of features we can extract for most modalities (e.g., acoustic or visual): knowledge-inspired features describing cues that were identified in psychological studies of human emotion recognition, and statistical features describing properties of the data.

For the model aspect, from a temporal view, models may use information from only the current time, or they can include contextual information; From a structural view, models may be flat using the input feature representations directly, or layered, designed to learn a better feature representation before performing classification or regression. Whether to choose one approach or the other, or to combine them, are questions faced by most emotion recognition researchers. In this work, we attempt to provide a better understanding of these issues by using the following as examples to compare these approaches:

- Data:
  - Spontaneous: the Audio/Visual Emotion Challenge 2012 (AVEC2012) database [22]
  - Acted: the Interactive Emotional dyadic MOtion CAPture (IEMOCAP) database [23]
- Features:
  - Knowledge-inspired: disfluencies and non-verbal vocalisations (DIS-NV) features [12]
  - Statistical: Low-Level Descriptors (LLD) features [24]
- Model:
  - Temporal:
    - Without context: Support Vector Machine (SVM) [25]
    - With context: Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) [26]
  - Structural:
    - Flat: SVM
    - Layered: LSTM-RNN

Our results show that the performance of different types of features and models largely depends on the nature of the data they are applied to. Thus, features and models should be chosen based on the type of dialogue and the amount of data that the task is performed on.

II. Databases and Related Work

We chose the AVEC2012 database of spontaneous dialogues and the IEMOCAP database of acted dialogues because they are the most widely used databases of English dialogues annotated with dimensional emotions.

A. The Dimensional Emotion Annotation

In this work, we use dimensional emotion annotation, in which emotions are represented as vectors in a multi-dimensional space. There are four commonly used emotional dimensions: Arousal, Expectancy, Power, and Valence [27]. The Arousal dimension describes the activeness of a subject; the Expectancy dimension describes whether the subject feels that the things under discussion are predictable (positive values) or surprising (negative values); the Power dimension describes whether the subject feels that (s)he dominates the conversation (positive values) or (s)he is being dominated (negative values). The Valence dimension describes whether the subject has positive feelings (positive values) or negative feelings (negative values) towards the topics under discussion.

B. The AVEC2012 Database

The AVEC2012 database [22] was collected as part of the SEMAINE corpus [28]. It includes approximately 8 hours of audio-visual recordings, auto generated word timings, and manually corrected and aligned transcripts of 24 subjects conversing with 4 on-screen characters role-played by human operators. Each character is designed with a different personality, namely even-tempered Prudence, happy Poppy, angry Spike, and depressive Obadiah. Topics of conversation vary from daily life to political issues, and each dialogue session is approximately five minutes long. The 24x4 recordings are divided into the training set, development set, and test set, each containing 32 dialogue sessions. In this work, we focus on word level emotion recognition, in which each word spoken by a subject is a data instance. The numbers of instances contained in the training set, development set, and test set are 20169, 16300, and 13405, respectively. We combined the training and development set for training our models (36469 training instances in total), and tested on the test set.

The AVEC2012 database uses real-value vectors in the Arousal-Expectancy-Power-Valence emotion space to represent emotions. It contains emotion annotations for each frame of the recording. The average of all the annotations was used when the annotators disagreed. The word level emotion annotations use the mean of all frames in a word as their emotion values. In this work, to get a clearer view of the relations between features and different emotions, we transformed each dimension of continuous emotion annotations into three discrete categories: low (value range [-1, -0.333]), medium (value range (-0.333, 0.333)), and high (value range [0.333, 1]).

C. The IEMOCAP Database

The IEMOCAP database [23] contains approximately 12 hours of audio-visual recordings from 5 mixed gender pairs of actors. The recordings were manually transcribed. Each conversation was approximately five minutes long. There are 10037 utterances in total, of which 4782 utterances were not scripted. When collecting the non-scripted dialogues, the actors were instructed to act out emotionally intense scenarios, e.g., telling a best friend that (s)he has been accepted into his/her most desired university. When collecting the scripted dialogues, the actors would follow pre-scripted lines to act out scenarios, such as an argument between a married couple.

Emotions were annotated at the utterance level with a 1 to 5 integer score of the Arousal, Power, and Valence emotion dimensions. The average of all the annotations was used when the annotators disagreed. We categorized the scores into three classes: low (with scores less than 3), medium (with scores equal to 3), and high (with scores larger than 3). This was done in order to have a clearer view of the relation between emotions and features, and to reduce the influence of imbalanced classes.

D. Related Work

Previous work on the AVEC2012 database has focused on using Low-Level Descriptor (LLD) based features for the acoustic model (e.g., [14], [29], [30]). However, there are results indicating that knowledge-inspired features, such as global prosodic features, may also be highly predictive (e.g., [15], [31]).

Most recognition models on the AVEC2012 database use Support Vector Regression without including contextual information. However, emotion is a relatively stable phenomenon and the emotional states of previous words are closely related to the current state. Therefore, the few models that have included contextual information, in either the features extracted [12] or the recognition model used (e.g., Hidden Markov Model [15], and Particle Filtering [29]), have shown better performance in emotion recognition. To the best of our knowledge, the only previous work on the AVEC2012 database that applies deep learning models used the LSTM-RNN model to learn better audio and visual feature representations separately, and then applied Support Vector Regression on the outputs of the LSTM-RNN models [32]. In their work, emotions were recognized at the frame level and their models were tested on the development set. Their results show that using LSTM-RNN models improves recognition performance compared to using the LLD features directly, which indicates that the LSTM-RNN model learned better feature representations.

In previous work on the IEMOCAP database, LLD features are also widely used for acoustic models (e.g., [33], [34]). However, recent work has shown that knowledge-inspired global prosodic features are more predictive than the LLD features for predicting binary Arousal values [35].

The LSTM-RNN model was used directly for classification in previous work on the IEMOCAP database. Results have shown that LSTM-RNN models have better performance than Hidden Markov Models (e.g., [36], [33]). Another application of deep learning methods uses Denoising Autoencoders to model gender information, which is shown to help with the emotion recognition task [37].

Because different settings were used in previous work, such as data preprocessing and focusing on different emotion annotations, it is hard to compare results. The different nature of emotion recognition tasks on the AVEC2012 database and the IEMOCAP database also means that results on these two databases are not directly comparable. Thus, in this work, we build our own models to compare the performance of knowledge-inspired and statistical acoustic features, when used with both a conventional classification model and a deep learning model. We performed experiments on both the AVEC2012 and the IEMOCAP databases.

III. Features and Models

In our experiments, we extracted the DIS-NV and the LLD features, and built the SVM and the LSTM-RNN models.

A. Features

1) DIS-NV Features: In this work, we study three types of disfluencies: filled pauses (non-verbal insertions, e.g., "eh"), fillers (verbal insertions, e.g., "you know"), and stutters (involuntarily repeats of part of a word or words); as well as two types of non-verbal vocalisations: laughter and audible breath. We chose them because they are the most common in the databases, and they are relatively easy to annotate from transcripts. We manually annotated DIS-NVs for both databases.

We used a moving window with a length of 15 words to compute the disfluency features for word-level emotion recognition on the AVEC2012 database. As shown in Figure 1, the window includes the current word and its 14 history words, and slides from the beginning of a dialogue session until its end. Feature values are calculated as the ratio between the sum duration of each type of DIS-NV appearing in the window and the total duration of the window (including silences between words). This results in 5 DIS-NV features for each word. We choose a window length of 15 words because this is the average length of an utterance. For utterance-level emotion recognition on the IEMOCAP database, we used the real utterances instead of the moving window.

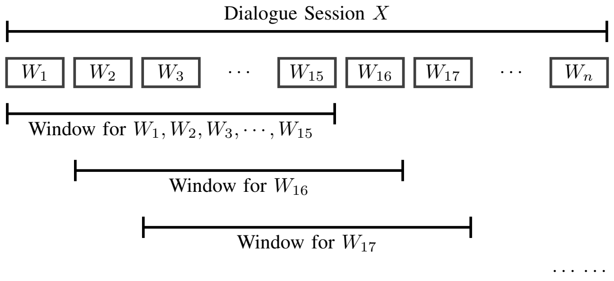

Fig. 1. Window for extracting DIS-NV features from the AVEC2012 database

TABLE I. FREQUENCIES OF DIS-NV

Databases	**FP**(%)	**FL**(%)	**ST**(%)	**LA**(%)	**BR**(%)
AVEC2012	32.0	14.7	9.4	11.9	2.7
IEMOCAP	11.2	24.1	6.3	1.6	0.6

Compared with the AVEC2012 database of spontaneous dialogues, DIS-NVs are less frequent in the IEMOCAP database of acted dialogues, which may limit their predictive power. 47.28% of the IEMOCAP non-scripted utterances and 24.74% of the IEMOCAP scripted utterances contain at least one type of DIS-NV. Table I shows the percentage of utterances containing each type of DIS-NV in both databases. In the first row, "FP" represents filled pause, "FL" is filler, "ST" is stutter, "LA" is laughter, "BR" is breath. Frequencies of most types of DIS-NV are lower on the IEMOCAP database. The fillers are the only exception, which may be because some fillers were part of the scripts. Because each pair of actors played out every script, fillers were duplicated when collecting the scripted dialogues of the IEMOCAP database.

2) LLD Acoustic Features: We extracted the LLD acoustic features by using a frame-level sliding window to transform the audio segment into a series of frames, then applying functionals (e.g., mean) to LLDs (e.g., MFCCs) and their corresponding delta coefficients. The OpenSMILE toolbox [24] was used to extract these features from audio recordings automatically for both databases. We use different LLD features for the AVEC2012 database and the IEMOCAP database. As we mentioned in Section II-D, results on the AVEC2012 and the IEMOCAP databases are not directly comparable, and it is difficult to compare with previous work because of differences in experimental settings. Thus, we chose the most widely used LLD feature set from previous work on each database as the reference set for experiments on this database. In the future, we will experiment with the union of the two LLD feature sets for both databases.

The LLD features used for the AVEC2012 database are the 1842 baseline audio features used in the AVEC2012 challenge. These features include 25 energy/spectral LLD × 42 functionals, 25 delta coefficients of the energy/spectral LLD × 19 functionals, 6 voicing related LLD × 32 functionals, 6 delta coefficients of the voicing related LLD × 19 functionals, and 10 voiced/unvoiced durational features.

The LLD features used for the IEMOCAP database are the 1582 baseline features used in the INTERSPEECH 2010 Paralinguistic Challenge. These features include 34 energy/spectral LLD × 21 functionals, 34 delta coefficients of the energy/spectral LLD × 21 functionals, 4 pitch-based LLD × 19 functionals, 4 delta coefficients of the pitch-based LLD × 19 functionals, and the number of pitch onsets (pseudo syllables) and the total duration of the input.

3) Summary of Features: We extracted two types of features in this work: knowledge-inspired DIS-NV features, which describe data at the utterance level with a small number of features; and statistical LLD features, which describe data at the frame level with a large feature set. The LLD features are able to give detailed information of all the data, while the DIS-NV features can highlight the utterances that may be specifically interesting for emotion recognition. The LLD feature sets have been widely used in current emotion recognition studies, while the DIS-NV features we proposed in our previous work [12] have shown performance improvement over other state-of-the-art unimodal feature sets in recognizing emotions in spontaneous dialogues on continuous scales.

B. *Classification Models*

We built two types of classification models in this work: a SVM model, which does not model sequence information and uses the given feature representations directly; and a LSTM-RNN model, which can automatically learn a flexible history length and an abstracted feature representation. Compared to the SVM model, the LSTM-RNN model has more parameters that need to be learned during training.

1) SVM: Our SVM models [25] were built with the LibSVM [38] classifier using WEKA [39]. We used the C-SVC approach with RBF kernel for both databases. All features were normalized to [-1,1] before classification. This is the setting widely used in previous work (e.g., [14], [22]).

2) LSTM-RNN: The LSTM-RNN model [26] is a neural network with multiple hidden layers and a special structure called "the memory cell" that can model long range context information. Compared to conventional RNN architectures, the LSTM-RNN model is able to learn from a longer history.

A hidden layer in a LSTM-RNN model is composed of recurrently connected memory blocks, each of which contains one or more recurrently connected memory cells. Each memory cell has three multiplicative "gate" units: the input, output, and forget gates. These gates perform the operations of reading, writing, and resetting, respectively. They allow the network to store and retrieve information over long periods of time. The structure of a LSTM memory cell is shown in Figure 2 [40]. "CEC" in the figure represents the "Constant Error Carousel", which is the central neuron that recycles status information from one time step to the next. The small blue circles with a cross inside indicate multiplicative connections. The peephole connection gives direct access to the central neuron.

We used the PyBrain toolbox [40] to build the LSTM-RNN models, with one memory cell in each memory block. For models on the AVEC2012 database, we used 16 memory cells for the DIS-NV feature set, 64 cells for the LLD feature set, and 64 cells for the DIS-NV and LLD concatenated feature set. For models on the IEMOCAP database, we used 16 cells for the DIS-NV feature set, 32 cells for the LLD feature set, and 32 cells for the concatenated feature set. All networks were trained using a learning rate of 10^{-5}. The number of memory

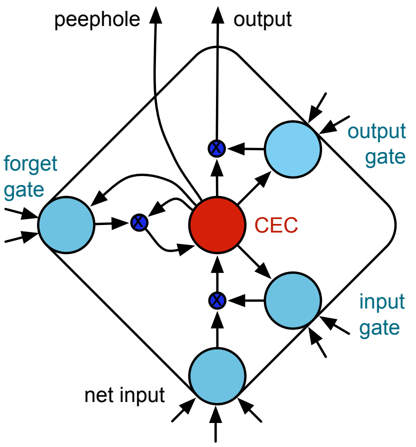

Fig. 2. Structure of a LSTM memory cell [40]

TABLE II. RESULTS ON THE AVEC2012 TEST SET

Features	Models	A (%)	E (%)	P (%)	V (%)	Mean
DN	SVM	52.4	61.4	67.4	59.2	60.1
	LSTM	**54.1**	**65.8**	**68.3**	**60.1**	**62.0**
LLD	SVM	52.4	60.8	67.5	59.2	60.0
	LSTM	52.4	60.7	66.1	58.1	59.3
DN+LLD	LSTM	52.5	61.2	65.8	58.0	59.4

cells was selected by cross-validation experiments, while the learning rate was selected following settings in [33].

IV. Results and Discussion

Results of our experiments on the AVEC2012 and the IEMOCAP databases are shown in Table II and Table III. The numbers are the weighted F-measures of the three classes of each emotion dimension, as defined in Section II-B. In the first row of the result tables, “A” represents Arousal; “E” is Expectancy; “P” is Power; “V” is Valence; “Mean” refers to the unweighted average of results on all emotion dimensions. In the first column, “DN” represents using the DIS-NV feature set alone; “LLD” is using the LLD acoustic feature set alone; “DN+LLD” is fusing the DIS-NV model and the LLD model at the feature-level, in which the DIS-NV feature set and the full LLD feature set are first concatenated, and then a LSTM-RNN model is applied to this concatenated feature set. Applying a SVM model to the concatenated feature set gave results that were not much different from the LLD-SVM models. Thus, we did not report these results in the tables.

A. Results on the AVEC2012 Database

Our results on the AVEC2012 database are shown in Table II. The models were trained on the AVEC2012 train+development set and tested on the unseen AVEC2012 test set. We used the train+development set to obtain more training instances for the LSTM-RNN models. We will perform cross-validation with the AVEC2012 database in the future. In SVM experiments with LLD features, Principal Components Analysis (PCA) preserving 95% of the total variance was applied to the LLD feature set to reduce dimensionality, resulting in 640 features. The LLD feature set used for LSTM-RNN models was the original full feature set.

When using SVM as the classification model, as in our previous work [12], performance of the DIS-NV features is at least as good as the performance of the LLD features, while reducing the number of features from 1842 to 5. This verifies the predictive power of the DIS-NVs for emotion recognition in spontaneous dialogues. The performance difference between the DIS-NV and the LLD features is less pronounced compared to our previous work with continuous emotion annotation [12]. This may be due to transforming the regression task to a classification task, and the use of different evaluation metrics.

TABLE III. RESULTS ON THE IEMOCAP DATABASE

Features	Models	A (%)	P (%)	V (%)	Mean
DN	SVM	36.3	40.7	32.8	36.6
	LSTM	41.6	37.8	34.0	37.8
LLD	SVM	**65.2**	**53.8**	**53.5**	**57.5**
	LSTM	53.7	46.2	38.6	46.2
DN+LLD	LSTM	53.9	51.6	39.5	48.3

Applying the LSTM-RNN model instead of the SVM model improves the performance of the DIS-NV features. However, for the LLD features, using the LSTM-RNN model instead of the SVM model does not give better results. This may be because the LSTM-RNN model for the LLD feature set has a more complex structure than the LSTM-RNN model for the DIS-NV features. The number of input neurons of a LSTM-RNN model equals the number of features. Thus, the LSTM-RNN model for the LLD feature set is much bigger and has many more parameters that need to be learned during training. We also find a large decrease in performance of the LLD-LSTM model on the test set compared to on the training set, which indicates that the LLD-LSTM model may have the issue of over-fitting.

Fusing the DIS-NV and the LLD models gives results close to the LLD-LSTM models. Feature-level fusion gives small improvements compared to the LLD model, but not compared to the DIS-NV model. The size difference between the DIS-NV and the LLD feature sets may be the reason for the limited improvements obtained by simply concatenating feature sets. The small number of DIS-NV features compared to the LLD features (5 V.S. 1842) may result in the network being dominated by the LLD features. The DIS-NV features and the LLD features also describe data at different levels. Thus, a better fusion strategy than combining the two feature sets at the same level is needed, such as adding the frame-level LLD features at the input layer of the network, and adding the utterance-level DIS-NV features at a higher level of the network structure.

B. Results on the IEMOCAP Database

Unlike the AVEC2012 database, the Expectancy emotional dimension was not annotated in the IEMOCAP database. Because the IEMOCAP database was not split into training and test partitions, we performed 10-fold cross-validation on the IEMOCAP database. The results are shown in Table III. We report averages of F-measure over the test set of each cross-validation experiment, in which the models were trained with 90% of the IEMOCAP database, and tested with the remaining 10% unseen data. This is standard for experiments with the IEMOCAP database.

Unlike our results on the AVEC2012 database, the DIS-NV features are less predictive than the LLD acoustic features on the IEMOCAP database. This may be caused by the DIS-NVs being less frequent in acted dialogues compared to spontaneous dialogues (as shown in Table I). This indicates that there are fundamental differences between spontaneous and acted dialogues, thus feature predictiveness is influenced by the type of dialogue, which is consistent with our previous work on the IEMOCAP database [13].

Applying the LSTM-RNN model instead of the SVM model improves performance of the DIS-NV features. However, similar to our results on the AVEC2012 test set, applying the LSTM-RNN model to the LLD features leads to a decrease in performance. The increase in performance of the DIS-NV features is smaller on the IEMOCAP database, while the decrease of the LLD features is more obvious. One reason may be that although the total lengths of recordings in the two databases are about the same, the different annotation levels (utterance vs. word) result in fewer training instances available for the LSTM-RNN model in the IEMOCAP database (9033 training instances for the IEMOCAP database, 36469 training instances for the AVEC2012 database). Thus, there may not be enough data to optimize the networks. Another reason may be that compared to word-level emotion recognition, utterance-level emotion recognition may benefit less from including long range context information. A similar result was found in the work of Metallinou et al. [34] on the IEMOCAP database, where a Hidden Markov Model gave better results than a LSTM-RNN model for recognizing the Arousal dimension of emotion at the utterance level using LLD features.

Similar to our results on the AVEC2012 database, using the DIS-NV and the LLD model at the feature-level gives better performance than using only the LLD features. This verifies that the DIS-NV features contain additional information to the low-level acoustic features.

C. Discussion

Our experiments on different types of features have shown that the knowledge-inspired DIS-NV features perform better than the statistical LLD features when recognizing emotions in spontaneous dialogues. In contrast, the DIS-NV features are less predictive than the LLD features in acted dialogue, which may be due to the infrequency of DIS-NV in acted dialogues. These findings reflect fundamental differences in spontaneous and acted speech.

Because of the complex structure of the LSTM-RNN model, it can only outperform the SVM model when there is enough training data available. The large number of parameters in the LSTM-RNN model may also lead to the problem of over-fitting and may not generalize well to unseen data. Thus, in future work, we will include more emotional databases of English dialogues with dimensional emotion annotations, and train the models on a merged dataset of available databases.

Our experiments on different types of models have shown that contextual information is useful for emotion recognition. However, we found that emotions may be more stable at the word level than at the utterance level. That is to say, the emotion of the current word is closely related to the emotions of other words within the same utterance, but the emotion of the current utterance may be less related to the emotions of other utterances of the speaker, especially those far away. This makes the LSTM-RNN model more suitable for word-level emotion recognition due to its ability to learn from long distance contexts. However, when performing utterance-level emotion recognition, long distance contexts are less useful, which makes the LSTM-RNN model less helpful. In the future, we will analyse the descriptive statistics of the emotion annotations on word level and utterance level to study the level of stability of emotions in dialogue.

V. Conclusion

In order to study the influence of types of dialogues on the performance of emotion recognition, we conducted experiments on the AVEC2012 database of spontaneous dialogues and the IEMOCAP database of acted dialogues. We extracted two types of acoustic features in this work: knowledge-inspired DIS-NV features, and statistical LLD features. We also compared SVM and LSTM-RNN as the classification models. Our results show that the performance of features and models is largely influenced by the dialogue type and the size of the data set.

Consistent with our previous work ([12], [13]), utterance-level knowledge-inspired features outperform frame-level statistical features when recognizing emotions in spontaneous dialogues, but not in acted dialogues. In this work, we find that LSTM-RNN models consistently give better performance than the SVM models when there is enough training data. However, the complex structure of a LSTM-RNN model limits its performance in emotional databases that have less training instances available, and may also lead to the problem of over-fitting. Including contextual information was shown to be helpful, although long distance contexts may be more useful when performing emotion recognition at the word or frame level.

We are aware that our comparisons are not yet complete. For example, we only compared flat models without context (the SVM models) and layered models with context (the LSTM-RNN models). We plan to fill in the gap by experimenting on flat models with context (e.g., Hidden Markov Models), and layered models without context (e.g., Neural Networks). In the future, we will study the predictive power of other knowledge-inspired acoustic features (e.g., global prosodic features). We will also work on building a hierarchical emotion recognition model that combines different types of features at different levels based on their nature, such as whether the features are utterance-level features or frame-level features. In the future, we will also examine whether our findings generalize to other databases of English dialogues annotated with dimensional emotion annotations, such as the Belfast naturalistic database [41].

Acknowledgment

We would like to thank Dr. Robin Lickley for his help in psycholinguistic studies of disfluencies, and Dr. Amos Storkey for his suggestions in deep learning methods. The first author, Leimin Tian, is funded by the Institute for Language, Cognition and Computation of the School of Informatics, the University of Edinburgh.

References

[1] R. W. Picard, *Affective computing*. MIT press, 2000.

[2] J. A. Singer and P. Salovey, *Remembered Self: Emotion and Memory in Personality*. Simon and Schuster, 2010.

[3] A. Bechara, "The role of emotion in decision-making: evidence from neurological patients with orbitofrontal damage," *Brain and cognition*, vol. 55, no. 1, pp. 30–40, 2004.

[4] S. Shott, "Emotion and social life: A symbolic interactionist analysis," *American journal of Sociology*, pp. 1317–1334, 1979.

[5] F. Pecune, M. Mancini, B. Biancardi, G. Varni, Y. Ding, and C. Pelachaud, "Laughing with a virtual agent," 2015.

[6] A. Popescu, J. Broekens, and M. van Someren, "Gamygdala: An emotion engine for games," *Affective Computing, IEEE Transactions on*, vol. 5, no. 1, pp. 32–44, 2014.

[7] J. Xu, J. Broekens, K. Hindriks, and M. A. Neerincx, "Effects of bodily mood expression of a robotic teacher on students," in *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*. IEEE, 2014, pp. 2614–2620.

[8] P. Vuilleumier, "How brains beware: neural mechanisms of emotional attention," *Trends in cognitive sciences*, vol. 9, no. 12, pp. 585–594, 2005.

[9] L. Vidrascu and L. Devillers, "Detection of real-life emotions in call centers." in *INTERSPEECH*, vol. 2005, no. 10, 2005, pp. 1841–1844.

[10] R. Lickley, "Fluency and disfluency," 2015, to appear.

[11] C. McGettigan, E. Walsh, R. Jessop, Z. Agnew, D. Sauter, J. Warren, and S. Scott, "Individual differences in laughter perception reveal roles for mentalizing and sensorimotor systems in the evaluation of emotional authenticity." *Cerebral cortex (New York, NY: 1991)*, vol. 25, no. 1, pp. 246–257, 2015.

[12] J. Moore, L. Tian, and C. Lai, "Word-level emotion recognition using high-level features," in *Computational Linguistics and Intelligent Text Processing*. Springer, 2014, pp. 17–31.

[13] L. Tian, C. Lai, and J. Moore, "Recognizing emotions in dialogues with disfluencies and non-verbal vocalisations," in *Proceedings of the 4th Interdisciplinary Workshop on Laughter and Other Non-verbal Vocalisations in Speech*, 2015.

[14] N. Lubis, S. Sakti, G. Neubig, T. Toda, A. Purwarianti, and S. Nakamura, "Emotion and its triggers in human spoken dialogue: Recognition and analysis," 2014.

[15] D. Ozkan, S. Scherer, and L.-P. Morency, "Step-wise emotion recognition using concatenated-HMM," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 477–484.

[16] T. Baltrusaitis, N. Banda, and P. Robinson, "Dimensional affect recognition using continuous conditional random fields," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[17] H. Kaya, F. Eyben, A. A. Salah, and B. Schuller, "Cca based feature selection with application to continuous depression recognition from acoustic speech features," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3729–3733.

[18] K. Forbes-Riley and D. Litman, "Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor," *Speech Communication*, vol. 53, no. 9, pp. 1115–1136, 2011.

[19] J. Schmidhuber, "Deep learning in neural networks: An overview," *CoRR*, vol. abs/1404.7828, 2014.

[20] R. Brueckner and B. Schuller, "Be at odds? deep and hierarchical neural networks for classification and regression of conflict in speech," in *Conflict and Multimodal Communication*. Springer, 2015, pp. 403–429.

[21] J. Deng, Z. Zhang, F. Eyben, and B. Schuller, "Autoencoder-based unsupervised domain adaptation for speech emotion recognition," 2014.

[22] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic, "AVEC 2012: the continuous audio/visual emotion challenge," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 449–456.

[23] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[24] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[25] C. Shawe-Taylor and S. Schölkopf, "The support vector machine," 2000.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[28] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*. IEEE, 2010, pp. 1079–1084.

[29] A. Savran, H. Cao, M. Shah, A. Nenkova, and R. Verma, "Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 485–492.

[30] A. Chen, S. Yuan, and D. Jiang, "Bagging based feature selection for dimensional affect recognition in the continuous emotion space," 2013.

[31] L. van der Maaten, "Audio-visual emotion challenge 2012: a simple approach," in *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 2012, pp. 473–476.

[32] J. Wei, E. Pei, D. Jiang, H. Sahli, L. Xie, and Z. Fu, "Multimodal continuous affect recognition based on lstm and multiple kernel learning," in *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*. IEEE, 2014, pp. 1–4.

[33] M. Wollmer, A. Metallinou, N. Katsamanis, B. Schuller, and S. Narayanan, "Analyzing the memory of blstm neural networks for enhanced emotion classification in dyadic spoken interactions," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4157–4160.

[34] A. Metallinou, M. Wollmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *Affective Computing, IEEE Transactions on*, vol. 3, no. 2, pp. 184–198, 2012.

[35] D. Bone, C. Lee, and S. Narayanan, "Robust unsupervised arousal rating: A rule-based framework with knowledge-inspired vocal features," *Affective Computing, IEEE Transactions on*, vol. 5, no. 2, pp. 201–213, 2014.

[36] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional lstm modeling." in *INTERSPEECH*, 2010, pp. 2362–2365.

[37] R. Xia, J. Deng, B. Schuller, and Y. Liu, "Modeling gender information for emotion recognition using denoising autoencoder," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 990–994.

[38] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[39] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[40] T. Schaul, J. Bayer, D. Wierstra, Y. Sun, M. Felder, F. Sehnke, T. Rückstieß, and J. Schmidhuber, "PyBrain," *Journal of Machine Learning Research*, 2010.

[41] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech communication*, vol. 40, no. 1, pp. 33–60, 2003.