RECOGNIZING EMOTIONS IN SPOKEN DIALOGUE WITH HIERARCHICALLY FUSED ACOUSTIC AND LEXICAL FEATURES

Leimin Tian, Johanna Moore, Catherine Lai

School of Informatics, the University of Edinburgh
Informatics Forum, 10 Crichton Street, Edinburgh, UK, EH8 9AB
s1219694@sms.ed.ac.uk, J.Moore@ed.ac.uk, clai@inf.ed.ac.uk

ABSTRACT

Automatic emotion recognition is vital for building natural and engaging human-computer interaction systems. Combining information from multiple modalities typically improves emotion recognition performance. In previous work, features from different modalities have generally been fused at the same level with two types of fusion strategies: Feature-Level fusion, which concatenates feature sets before recognition; and Decision-Level fusion, which makes the final decision based on outputs of the unimodal models. However, different features may describe data at different time scales or have different levels of abstraction. Cognitive Science research also indicates that when perceiving emotions, humans use information from different modalities at different cognitive levels and time steps. Therefore, we propose a Hierarchical fusion strategy for multimodal emotion recognition, which incorporates global or more abstract features at higher levels of its knowledge-inspired structure.

We build multimodal emotion recognition models combining state-of-the-art acoustic and lexical features to study the performance of the proposed Hierarchical fusion. Experiments on two emotion databases of spoken dialogue show that this fusion strategy consistently outperforms both Feature-Level and Decision-Level fusion. The multimodal emotion recognition models using the Hierarchical fusion strategy achieved state-of-the-art performance on recognizing emotions in both spontaneous and acted dialogue.

***Index Terms*—** emotion recognition, modality fusion, LSTM, dialogue, human-computer interaction

1. INTRODUCTION

Cognitive science research has shown that emotions are vital in human cognition and communication processes [1]. It has become increasingly apparent that recognizing emotions in spoken dialogue is crucial for advancing human-computer interaction technologies. For example, a virtual agent able to copy and adapt its laughter and expressive behaviour to user's behaviours has been shown to increase the user's humour experience [2]. Similarly, in a teaching scenario, a robot lecturer monitoring the emotional states of the students and expressing a positive mood while giving lectures to university classes was rated as having higher lecturing quality [3]. This has led to growing interest in automatic emotion recognition.

Similar to human emotion recognition, combining multiple modalities typically improves automatic emotion recognition performance. However, the improvement is often limited [4]. One reason may be that previous multimodal models combine modalities at the same level. That is, unimodal models are either combined at the Feature Level by concatenating feature sets (FL fusion), or at the Decision Level by fusing predictions made by each unimodal model (DL fusion). However, different modalities may describe data at different time scales or levels of abstraction. For example, many wavelet based acoustic features describe data at the frame level, while many lexical features describe data at the word or utterance level. Statistical features may describe detailed information unrelated to emotions, while knowledge-inspired features are generally more abstract and describe emotion-specific cues. Moreover, Cognitive Science studies also indicate that when perceiving emotions, humans make use of information at different cognitive levels and time steps [5].

The above observations indicate that compared to modality fusion at the same level, modality fusion at different levels may increase the benefits gained. Moreover, in FL fusion, features are often concatenated without awareness of modality differences, while in DL fusion, detailed information about the features within each modality is lost in the decision-making model. Therefore, we are motivated to develop a HierarchicaL (HL) fusion strategy, which incorporates features at different levels of its knowledge-inspired structure: features that describe data at larger time scales and are more abstract are used in higher levels. Compared to FL fusion, the knowledge-inspired structure of HL fusion allows us to implement prior knowledge of modality differences. Compared to DL fusion, HL fusion is able to preserve detailed information when making the decision, and to incorporate feature differences within a modality. Therefore, our hypothesis is that HL fusion will achieve better performance than FL or DL fusion for multimodal emotion recognition.

To study the performance of this HL fusion strategy, we extract six state-of-the-art acoustic and lexical feature sets, and build multimodal emotion recognition models with HL, FL, and DL fusion strategies, respectively. We perform experiments on two widely used emotion databases of spoken dialogue: the AVEC2012 [6] and the IEMOCAP [7] databases. Our experiments show that HL fusion consistently outperforms FL or DL fusion on predicting all emotion dimensions on both databases. This indicates that our HL fusion strategy is useful for multimodal emotion recognition.

2. BACKGROUND AND RELATED WORK

2.1. Definition of Emotions

How to define emotions remains an open question in Cognitive Science. Many automatic emotion recognition studies follow the Darwinian emotion theory, which defines emotions in terms of several primary and universal categories, such as Ekman's Big-6 emotion categorization [8]. However, our work employs the cognitive emotion theory. This theory associates emotions with specific appraisals (stimuli that evoke changes in emotional states) and defines emotions as vectors in a space determined by a set of primitive emotion dimensions. The reason for using this theory is that most current emotional interaction modules in dialogue systems have been developed with appraisal-based emotion models, and our goal is to build emotion recognition models that can potentially be applied to such systems. In this work, we use four emotion dimensions that have been identified as being able to describe most everyday human emotions [9]: Arousal (activeness), Expectancy (predictiveness), Power (dominance), and Valence (positive/negative).

2.2. Automatic Emotion Recognition

Current automatic emotion recognition studies focus on identifying predictive feature representations and model structures. In this section, we review state-of-the-art approaches for emotion recognition.

2.2.1. *Features*

Features used for emotion recognition can be extracted from various modalities (e.g., audio, visual, physiological). Our work focuses on features extracted from the acoustic and lexical modalities because our task is to recognize emotions from spoken dialogue.

For the acoustic modality, previous studies have focused on Low-Level Descriptor (LLD) based features (e.g., [10, 11]), which describe statistical characteristics of the speech signal at the frame level. However, studies have shown that knowledge-inspired acoustic features describing utterance level prosodic patterns may have comparable or better performance than LLD features, while greatly reducing feature dimensionality (e.g., [12, 13]).

For the lexical modality, sparse features drawn from hand crafted affective dictionaries are dominant in current studies, e.g., Linguistic Inquiry and Word Count (LIWC) [14] based lexical features [15] and WordNetAffect [16] based lexical features [17]. However, current Paralinguistic studies on human-human dialogue suggest that besides lexical content, other phenomena in speech are also indicators of emotion. For example, disfluency conveys information such as uncertainty of the speaker [18]. Non-verbal vocalisations, especially laughter, have also been identified as universal and basic cues in human emotion recognition [19]. In our previous work [20, 21], features describing occurrences of disfluencies and non-verbal vocalisations in utterances have been shown to be predictive for recognizing emotions in spontaneous dialogue.

Besides identifying predictive features, feature engineering is also important for developing accurate emotion recognition models. In previous studies, applying feature engineering methods such as Canonical Correlation Analysis [22] or Correlation-based Feature-subset Selection [20] to the extracted features has been shown to improve emotion recognition performance.

2.2.2. *Models*

Beyond features, emotion recognition performance also depends on how features are modelled. Most widely used machine learning algorithms have been applied to emotion recognition. For example, Support Vector Machines [10] and Naive Bayes models [23]. However, previous work comparing various shallow learning algorithms (e.g., Support Vector Machines) indicates that performance differences are not significant after controlling for feature sets and other parameter settings [24].

Recently, deep learning models have improved emotion recognition performance compared to shallow learning algorithms (e.g., [25, 26]). The network structure of deep learning models allows auto abstraction of feature representations. Among different deep learning models, emotion recognition researchers are particularly interested in the Long Short-Term Memory Recurrent Neural Network (LSTM) because of its ability to model long-range contexts (e.g., [27]). The Bidirectional-LSTM, a modification of LSTM that includes context from both the past and the future, is also popular in state-of-the-art emotion recognition studies (e.g., [28]). However, it has been argued that Bidirectional-LSTMs do not improve performance significantly compared to using standard LSTMs [29]. To achieve better results, previous work has focused on identifying more predictive feature representations to input to the LSTM model (e.g., [30]) and stacking other machine learning models on top of the LSTM model (e.g., [31]).

One issue with using deep learning models for emotion recognition is that emotions are expensive to annotate, thus emotion databases are often small in size, which can limit optimization of the complex deep learning models. Using a small set of knowledge-inspired features often results in better performance than using a large set of noisy statistical features (e.g., [21, 32]). It is also important to identify an effective structure for combining feature sets representing different levels of abstraction.

2.3. Multimodal Emotion Recognition

Humans convey and perceive emotions through all communicative modalities. Emotion recognition performance of a human typically improves when information from multiple modalities is available [33]. Consistent with human studies, multimodal emotion recognition models typically outperform unimodal models [4]. There are two main types of fusion strategy used in current multimodal emotion recognition: Feature-Level (FL) fusion (or "early fusion") and Decision-Level (DL) fusion (or "late fusion").

In FL fusion (e.g., [15]), feature sets from different modalities are concatenated before performing recognition, as shown in Figure 1. In some studies, feature engineering is first applied to the concatenated feature set or individual feature sets (e.g., [17]). However, it is hard to apply knowledge about different modalities in FL fusion. In contrast, DL fusion applies a rule-based decision model (e.g. [34]) or a machine learning model (e.g. [27]) over the outputs given by each unimodal model, as shown in Figure 2. Previous studies comparing these two fusion strategies show that DL fusion often outperforms FL fusion [11, 35]. However, detailed information about features within each modality is lost in the final decision model of DL fusion.

Both FL and DL fusion incorporate modalities at the same level. However, as discussed in Section 2.2.1, different features may describe data at different time scales or levels of abstraction. An empirical study on relations between emotions and verbal and non-verbal behaviours in human communication, such as speech audio, facial expression, and gesture, has suggested that temporal relationships between different modalities convey information related to emotions [36]. Cognitive studies also indicate that humans perceive speech in a multi-level manner [5]. The cognitive process during human dialogue is often defined as a four-level structure [37]. For example, the communication model proposed by Clark [38] characterizes communication into four steps: attention, identification, understanding and consideration. At the attention step, the listener becomes aware that his/her conversational partner is speaking. At the identification step, the listener perceives the acoustic variations and recognizes the content of the speech (s)he hears. After identifying the content, in the understanding step, the listener analyses the meaning of the sentences (s)he just recognized. At the final consideration step, the meaning conveyed in the perceived speech evokes specific reactions and verbal/emotional responses of the listener based on his/her personal memories, knowledge, or goals. Under such a model, acoustic features are perceived earlier than the lexical features and at a lower cognitive level.

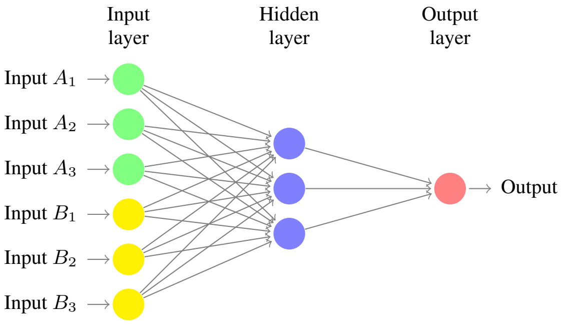

Fig. 1. An example of Feature-Level (FL) fusion model.

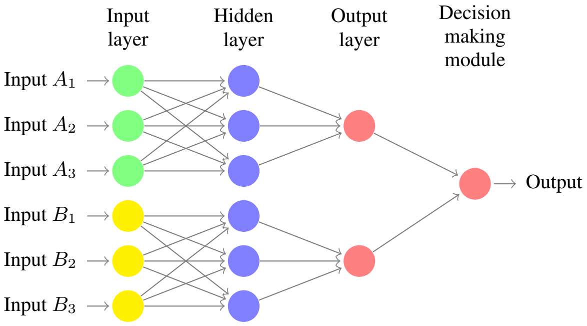

Fig. 2. An example of Decision-Level (DL) fusion model.

3. HIERARCHICAL FUSION

To address the limitations of FL and DL fusion, we propose a HierarchicaL (HL) fusion strategy, which incorporates features that describe data at a larger time scale or are more abstract at higher levels of its hierarchical structure, as shown in Figure 3.

To the best of our knowledge, the only previous work using a similar hierarchical approach for multimodal emotion recognition is by Chen et al [29]. In their work, features from the audio, visual, and physiological modalities were used to recognize frame level continuous Arousal and Valence values in French dyadic dialogue. Chen's hierarchical model differentiates between modalities, but does not take differences between features within a single modality into account. However, in a previous work on emotion recognition with frame level statistical acoustic features [39], a logistic regression model incorporating features derived from prosody, spectral envelope, and glottal information in a hierarchical structure outperformed a logistic regression model using all acoustic features at the input level. This indicates that a hierarchy capturing differences both between and within modalities is desirable for multimodal emotion recognition. The motivation of Chen's hierarchical fusion is to address the fact that signals from different modalities change asynchronously. Chen's hierarchical fusion outperformed FL fusion, but performed worse than DL fusion. The fact that different feature sets may have different levels of abstraction may have limited the performance of Chen's model.

Compared to Chen's hierarchical model, our HL model has a knowledge-inspired structure that incorporates both inter- and intra-modality differences. The hierarchy of our HL model is motivated both by the temporal characteristics and the levels of abstraction of the features. In the following sections, we present experiments that show our HL fusion can outperform both FL and DL fusion.

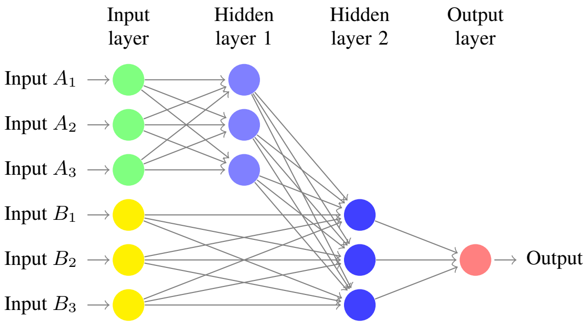

Fig. 3. An example of our HierarchicaL (HL) Fusion model.

4. METHODOLOGY

4.1. Emotion Databases of English Dialogue

Our experiments were conducted on the Audio/Visual Emotion Challenge 2012 (AVEC2012) database [6] and the IEMOCAP database [7]. They are the most widely used databases of English dialogues annotated with dimensional emotions.

The AVEC2012 database of spontaneous dialogue contains the Solid-SAL section of the SEMAINE corpus [40]. It includes approximately 8 hours of audio-visual recordings and manual transcripts of 24 subjects conversing with 4 on-screen characters with specific personalities role-played by human operators. Emotions were annotated at the word level as continuous values on the Arousal, Expectancy, Power, and Valence dimensions.

The IEMOCAP database of acted dialogue contains approximately 12 hours of audio-visual recordings from 5 mixed gender pairs of actors. There are two types of dialogue in the IEMOCAP database: non-scripted and scripted dialogue. When collecting the non-scripted dialogue, the actors were instructed to act out scenarios (e.g., customer service) without a pre-written script. When collecting the scripted dialogue, the actors followed pre-written lines. Emotions were annotated at the utterance level as a 1 to 5 integer score on the Arousal, Power, and Valence dimensions.

In order to unite different annotations of these two databases and to address the class imbalance issue, we transformed the original annotations of both databases into three discrete classes (*low*, *medium*, and *high*) on each emotion dimension.

It is hard to compare our results directly with previous studies conducted on these two databases, because previous results were achieved under various experimental settings and report different evaluation metrics. Thus, we extract state-of-the-art features and build emotion recognizers with different modelling approaches using these features to study the efficacy of the HL fusion strategy compared to state-of-the-art emotion recognition approaches.

4.2. Acoustic and Lexical Features

In this work, we extracted six state-of-the-art acoustic and lexical feature sets. Z-score speaker normalization was applied to all the features before performing emotion recognition.

4.2.1. *Acoustic Features*

LLD Features: The LLD features are statistical features extracted using a frame level sliding window of 25ms. Functionals (e.g., mean) were applied to LLDs (e.g., MFCCs) and their corresponding delta coefficients. The OpenSMILE toolbox [41] was used to automatically extract these features from audio recordings. We chose the InterSpeech 2010 Paralinguistic Challenge feature set in this work because this is a widely used benchmark set for emotion recognition. This feature set contains 1582 LLD features: 21 functionals applied to 34 LLDs with their corresponding delta coefficients, 19 functionals applied to 4 pitch-based LLDs and their corresponding delta coefficients, the number of pitch onsets (pseudo syllables) and the total duration of the input. A list of the functionals and LLDs can be found in Section 2.5.5 of [42].

eGeMAPS Features: The eGeMAPS features are frame-level, knowledge-inspired features. The set contains LLD features that haven been suggested as the most related to emotions by Paralinguistic studies [43]. The OpenSMILE toolbox was used again to extract these features automatically. Cross-corpora studies indicate that these features have comparable or better performance than large LLD feature sets while greatly reducing the feature dimensionality [43]. The eGeMAPS feature set contains 88 features: the arithmetic mean and coefficient of variation of 18 LLDs, 8 functionals applied to pitch and loudness, 4 statistics over the unvoiced segments, 6 temporal features, and 26 additional cepstral parameters and dynamic parameters. A list of the functionals and LLDs can be found in Section 3 of [43]. The eGeMAPS feature set is a widely used benchmark set for emotion recognition studies.

Global Prosodic (GP) Features: The GP features are utterance-level, knowledge-inspired prosodic features based on the work of Bone et al [12]. These include three features: median pitch, median intensity, and voice quality (HF500) over the utterance. HF500 is a spectral-slope measurement computed as the ratio between the total energy above and below 500Hz in an utterance. These features were highly predictive of Arousal in previous work [12].

4.2.2. *Lexical Features*

Disfluency and Non-verbal Vocalisation (DIS-NV) Features: The five DIS-NV features are utterance-level, knowledge-inspired features based on manual annotations of three types of disfluency and two types of non-verbal vocalisation. This includes filled pauses (non-verbal insertions, e.g., "Hmm" in "Hmm, that's interesting."), fillers (verbal insertions, e.g., "you know" in "I just want to, you know, get a drink and forget all about it."), stutters (involuntary repeat of words or part of a word, e.g., "Sa" in "Sa... Saturday will be fine."), laughter and audible breath [20]. The feature values are the ratios between the durations of DIS-NV events and the utterance duration. Compared to the AVEC2012 database of spontaneous dialogue, DIS-NVs are less frequent in the IEMOCAP database of acted dialogue, which limits their performance on emotion recognition in acted dialogue [21, 44].

Point-wise Mutual Information (PMI) Features: PMI is a widely used measurement of the relation between words and emotions. It is based on the frequency of a word being annotated as an emotion. PMI features are utterance-level, knowledge-inspired features and have been shown to be highly predictive of emotions in previous work [20, 45]. Feature values are calculated as the sum of PMI values of all the words in an utterance for each binarized emotion dimension [20]. Eight PMI features were extracted for the AVEC2012 database. Six PMI features were extracted for the IEMOCAP database because only three emotion dimensions were annotated in this database.

Crowd-Sourced Emotion Annotation (CSA) Features: Because the PMI features are calculated for specific database that the emotion recognition task is performed on, they may not generalize well to unseen data. Therefore, we also extract 63 utterance-level, knowledge-inspired features based on crowd-sourced annotations of Arousal, Power, and Valence of approximately 14,000 English lemmas [46]. The reason we chose this dictionary resource is because it is an expanded dictionary compared to the affective dictionaries used in previous studies.

4.3. The LSTM Model

We build LSTM models for all emotion recognizers in this work. The LSTM model is a recurrent neural network with multiple hidden layers and a special structure called "the memory cell" that can model long-range context information. As shown in Figure 4, 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. In Figure 4, "CEC" 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. The reason we chose the LSTM model is because emotion is context dependent, thus the memory cell structure of a LSTM model is especially useful for emotion recognition tasks. As discussed in Section 2.2.2, the LSTM model has achieved leading performance in current emotion recognition studies. Our previous work comparing different features and recognition models for emotion recognition on the AVEC2012 and the IEMOCAP databases has also shown that the LSTM model generally achieves better performance than the widely used Support Vector Machines [21].

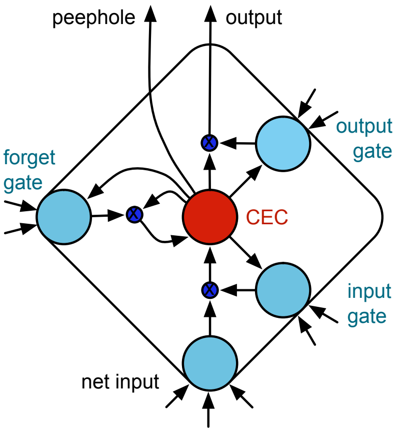

Fig. 4. Structure of a LSTM memory cell [47].

We optimize the model with 10-fold cross-validation experiments. LSTM models with a single hidden layer are used for the unimodal models and the final decision model of the DL model.[1]

[1] Number of memory cells: $LLD = 32$, $CSA = 16$, $GP = DN = PMI = 8$; $DL_{all} = 16$, $DL_{sub} = 8$.

The number of neurons in the input layer equals the total number of features used in this LSTM model. Our LSTM models are built with the PyBrain toolbox [47]. We used the R-Propagation-Minus trainer with a learning rate of 10^{-5}. All training samples have the same weight. An early stopping strategy was used to prevent over-fitting.

4.4. HL Fusion and Multimodal Models

As discussed in Section 2.3 and Section 3, we propose HL fusion to overcome limitations of FL and DL fusion. For multimodal models using all acoustic and lexical features, we used a LSTM model with three hidden layers in the FL and HL fusion models.[2] The reason we use the same LSTM model structure in the HL and the FL models is to remove other factors influencing the performance, such as the representation power of the hidden layers. The FL fusion model used all of the features as inputs to the bottom hidden layer.

For HL fusion, the LLD and eGeMAPS features are used as inputs to the bottom hidden layer, the GP and DIS-NV features are added at the middle hidden layer, and the PMI and CSA features are added at the top hidden layer, as shown in Figure 5. The LLD and eGeMAPS features are used at the bottom level because they are frame-level features, while the other features are utterance-level features. The PMI and CSA features are used at the top level because they encode prior information about emotional states, thus have a higher level of abstraction.

For multimodal emotion recognition models using only subsets of all features, we simplify the LSTM model by removing the input neurons connected to the removed features. Structures of the HL models using subsets of all features are shown in Figure 6 and Figure 7. Same with the HL models, we use a LSTM model with two hidden layers for the FL model using a subset of all features on the AVEC2012 database, and a LSTM model with three hidden layers for the FL model on the IEMOCAP database. Note that the number of neurons shown in these figures is only an indication and is not the number of neurons used in the real models.

For DL fusion, predictions given by unimodal models are used as inputs to another LSTM model. We also tested using rule-based decision models (e.g., select the class with highest confidence). However, these models performed worse than DL fusion using a LSTM model as the decision model, thus we focus on the later.

5. RESULTS AND DISCUSSION

Results of 10-fold cross-validation experiments on the two databases are shown in Table 1 and Table 2. We report weighted F-measures to address the class imbalance issue. In both tables, “A” is Arousal, “E” is Expectancy, “P” is Power, “V” is Valence. “Mean” is the arithmetic mean over all emotion dimensions. Note that the IEMOCAP database does not provide Expectancy annotations. Thus, results on the Expectancy dimension are missing for the IEMOCAP results.

For unimodal results, consistent with previous studies [12, 21, 43, 45], the knowledge-inspired features perform better than the statistical LLD features on both spontaneous and acted dialogue. This indicates that the ability of deep learning models to perform automatic feature abstraction has room for improvement. When the amount of training data is limited, using more predictive, knowledge-inspired features will result in better emotion recognition performance than using noisy statistical features. Thus, identifying predictive features remains an effective way to improve state-of-the-art of emotion recognition. The results also show that the predic-

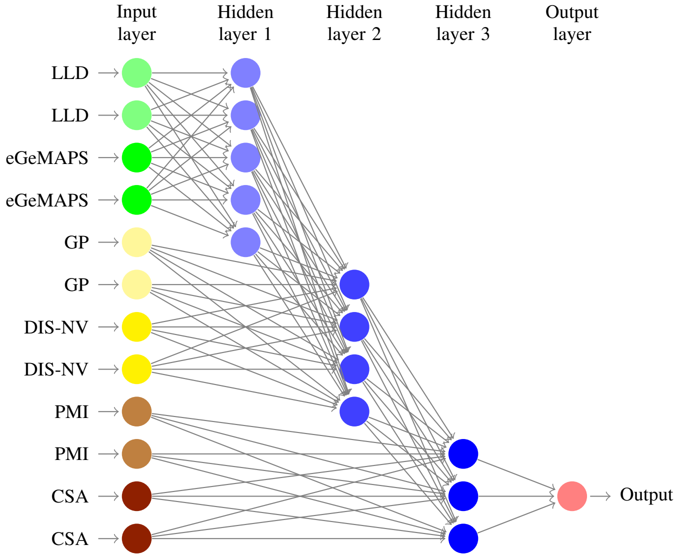

Fig. 5. Structure of HL fusion model using all features.

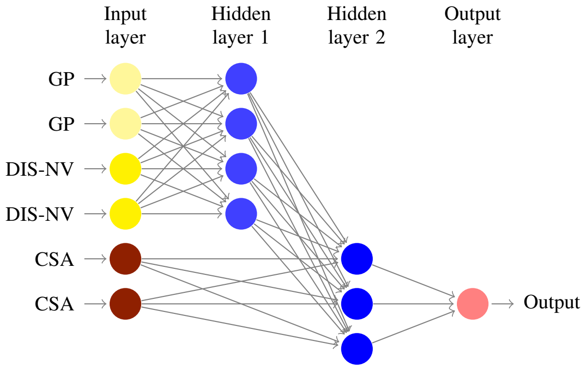

Fig. 6. Structure of the HL fusion model for the AVEC2012 database using only GP, DIS-NV and CSA features.

Table 1. Results on the AVEC2012 Database

Models	A(%)	E(%)	P(%)	V(%)	Mean(%)
Unimodal LSTM Models					
LLD	56.5	61.6	72.1	66.4	64.2
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**
Multimodal LSTM Models					
Combining all feature sets					
FL	56.4	63.4	72.6	67.9	65.1
DL	58.7	65.2	73.4	69.1	66.6
HL	**59.2**	**67.8**	**73.6**	**70.7**	**67.9**
Global Prosody + DIS-NV + CSA					
FL	60.1	68.1	74.8	71.7	68.7
DL	56.6	63.3	73.5	68.0	65.3
HL	**61.8**	**69.2**	**76.2**	**72.4**	**69.9**

[2] Number of memory cells: $h_{bottom} = 32$, $h_{middle} = 16$, $h_{top} = 8$.

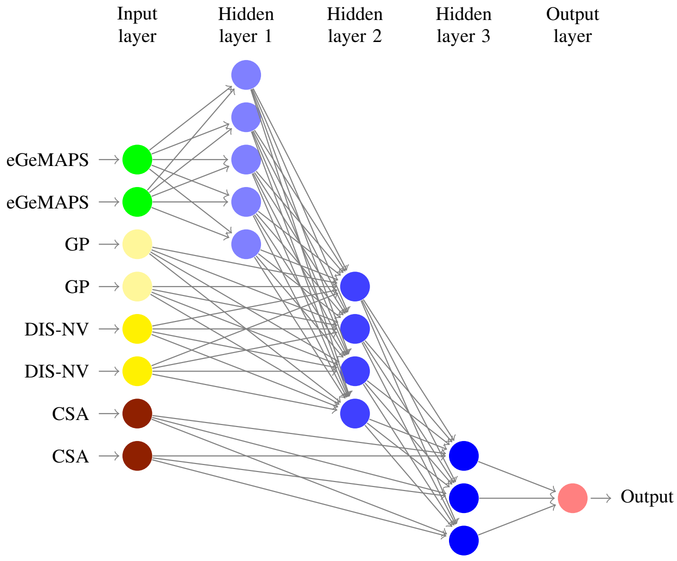

Fig. 7. Structure of the HL fusion model for the IEMOCAP database using only eGeMAPS, GP, DIS-NV and CSA features.

Table 2. Results on the IEMOCAP Database

Models	A(%)	P(%)	V(%)	Mean(%)
Unimodal LSTM Models				
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
Multimodal LSTM Models				
Combining all feature sets				
FL	53.4	48.7	37.1	46.4
DL	52.4	50.3	**47.4**	50.0
HL	**57.3**	**51.1**	*45.4*	**51.3**
eGeMAPS + Global Prosody + DIS-NV + CSA				
FL	55.2	50.8	47.2	51.1
DL	51.6	49.7	46.8	49.3
HL	**61.7**	**52.8**	**51.2**	**55.3**

tiveness of features depends on the specific task. As shown in the tables, acoustic features are more predictive on the acted IEMOCAP database than on the spontaneous AVEC2012 database. For example, the eGeMAPS features are the most predictive feature set on the IEMOCAP database, but they are less predictive than the lexical features on the AVEC2012 database. This is consistent with previous findings that in general emotions are acoustically exaggerated under acting scenarios [33].

For multimodal models using all feature sets, HL fusion outperforms FL fusion on all emotion dimensions on both databases. In addition, HL fusion outperforms DL fusion on most emotion dimensions on both databases. The only exception is the Valence dimension of the IEMOCAP database. As shown in Table 2, the large LLD feature set is not highly predictive on the Valence dimension for the IEMOCAP database. As the negative influence of the LLD feature set is larger for the HL model than for the DL model, the HL model performs worse than the DL model in this particular case. Moreover, recall that emotions were annotated at the utterance level on the IEMOCAP database, while on the AVEC2012 database emotions were annotated at the word level. Thus there are fewer training instances in the IEMOCAP database (approximately 10,000) than in the AVEC2012 database (approximately 50,000). This limits the performance of the HL model, which has more parameters to fit than the DL model. The smaller number of training samples in the IEMOCAP database compared to the AVEC2012 database may also be the reason that performance of these emotion recognizers using deep learning models is worse on the acted IEMOCAP database than on the spontaneous AVEC2012 database.

To improve performance of the multimodal emotion recognition models, we build multimodal models using only smaller, knowledge-inspired feature sets. We remove the PMI features for both databases because they are redundant with respect to the more general CSA features. For the AVEC2012 database, we remove both the LLD and eGeMAPS features because they are relatively large in size yet low in predictiveness. For the IEMOCAP database, we only remove the LLD features because the eGeMAPS features are highly predictive on the IEMOCAP database.

Results of the improved multimodal models are reported in Table 1 and Table 2. As we can see, performance of HL and FL models increases when using only smaller, knowledge-inspired feature sets. Performance of DL models decreases compared to using all features because the decision making module has less input information. When using only these knowledge-inspired features, HL fusion outperforms both FL and DL fusion on predicting all emotion dimensions on both databases. The HL model also achieves the best performance on all emotion dimensions on both databases compared to all the other unimodal and multimodal emotion recognition models. Compared to Chen's hierarchical model [29], our HL model has a knowledge-inspired structure that incorporates differences both between and within modalities, and achieves results better than both FL and DL fusion. This verified the efficacy of our HL model on multimodal emotion recognition in spoken dialogue.

6. CONCLUSIONS

We proposed a HierarchicaL (HL) fusion strategy for multimodal emotion recognition, which incorporates features in a knowledge-inspired hierarchy. We compared HL fusion with Feature-Level (FL) and Decision-Level (DL) fusion. Experiments on two emotion databases of spoken dialogue show that HL fusion consistently outperforms FL and DL fusion. The HL model achieves state-of-the-art performance for recognizing emotions in spoken dialogue, and has the potential to be applied to dialogue systems to improve the quality of emotional interaction. However, lack of training data may limit the performance of the HL emotion recognition model. Thus, we plan to study semi-supervised approaches for multimodal emotion recognition using HL fusion in the future.

While designed for emotion recognition, HL fusion could in principle, be applied to other multimodal recognition tasks, as it allows us to incorporate specific knowledge of feature abstraction and time scales.

7. ACKNOWLEDGEMENTS

We would like to thank Dr. Amos Storkey for his advice on deep learning. The first author, Leimin Tian, is funded by the Institute for Language, Cognition and Computation (ILCC) of the School of Informatics, the University of Edinburgh.

8. REFERENCES

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

[2] Florian Pecune, Maurizio Mancini, Beatrice Biancardi, Giovanna Varni, Yu Ding, and Catherine Pelachaud, "Laughing with a virtual agent," pp. 1817–1818, 2015.

[3] Junchao Xu, Joost Broekens, Koen Hindriks, and Mark A Neerincx, "Effects of bodily mood expression of a robotic teacher on students," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2014, pp. 2614–2620.

[4] Sidney D'Mello and Jacqueline Kory, "Consistent but modest: a meta-analysis on unimodal and multimodal affect detection accuracies from 30 studies," in *Proceedings of ACM-ICMI 2012*. ACM, 2012, pp. 31–38.

[5] Didier Grandjean, David Sander, and Klaus R Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and cognition*, vol. 17, no. 2, pp. 484–495, 2008.

[6] Björn Schuller, Michel Valster, Florian Eyben, Roddy Cowie, and Maja Pantic, "AVEC 2012: the continuous audio/visual emotion challenge," in *Proceedings of ACM-ICMI 2012*. ACM, 2012, pp. 449–456.

[7] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[8] Paul Ekman, Wallace V Friesen, Maureen O'Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause, William Ayhan LeCompte, Tom Pitcairn, Pio E Ricci-Bitti, et al., "Universals and cultural differences in the judgments of facial expressions of emotion.," *Journal of personality and social psychology*, vol. 53, no. 4, pp. 712, 1987.

[9] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth, "The world of emotions is not two-dimensional," *Psychological science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[10] Chung-Hsien Wu, Wei-Bin Liang, Kuan-Chun Cheng, and Jen-Chun Lin, "Hierarchical modeling of temporal course in emotional expression for speech emotion recognition," in *Proceedings of ACII 2015*. IEEE, 2015, pp. 810–814.

[11] Lang He, Dongmei Jiang, and Hichem Sahli, "Multimodal depression recognition with dynamic visual and audio cues," in *Proceedings of ACII 2015*. IEEE, 2015, pp. 260–266.

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

[13] Iulia Lefter, Harold T Nefs, Catholijn M Jonker, and Leon JM Rothkrantz, "Cross-corpus analysis for acoustic recognition of negative interactions," in *Proceedings of ACII 2015*. IEEE, 2015, pp. 132–138.

[14] JW Pennebaker, CK Chung, M Ireland, A Gonzales, and RJ Booth, "Operator's manual: Linguistic inquiry and word count," *LIWC2007. Austin, TX: LIWC*, 2007.

[15] Zahra Nazari, Gale Lucas, and Jonathan Gratch, "Multimodal approach for automatic recognition of machiavellianism," in *Proceedings of ACII 2015*. IEEE, 2015, pp. 215–221.

[16] Carlo Strapparava, Alessandro Valitutti, et al., "Wordnet affect: an affective extension of wordnet.," in *LREC*, 2004, vol. 4, pp. 1083–1086.

[17] Sonja Gievska, Kiril Koroveshovski, and Natasha Tagasovska, "Bimodal feature-based fusion for real-time emotion recognition in a mobile context," in *Proceedings of the ACII 2015*. IEEE, 2015, pp. 401–407.

[18] RoBin J LickLey, "Fluency and disfluency," *The handbook of speech production*, p. 445, 2015.

[19] C McGettigan, E Walsh, R Jessop, ZK Agnew, DA Sauter, JE Warren, and SK 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.

[20] Johanna Moore, Leimin Tian, and Catherine Lai, "Word-level emotion recognition using high-level features," in *Computational Linguistics and Intelligent Text Processing*. 2014, pp. 17–31, Springer.

[21] Leimin Tian, Johanna D Moore, and Catherine Lai, "Emotion recognition in spontaneous and acted dialogues," in *Proceedings of ACII 2015*. IEEE, 2015, pp. 698–704.

[22] Heysem Kaya, Florian Eyben, Albert Ali Salah, and Bjorn Schuller, "CCA based feature selection with application to continuous depression recognition from acoustic speech features," in *Proceedings of ICASSP 2014*. IEEE, 2014, pp. 3729–3733.

[23] Florian B Pokorny, Franz Graf, Franz Pernkopf, and Bjorn W Schuller, "Detection of negative emotions in speech signals using bags-of-audio-words," in *Proceedings of ACII 2015*. IEEE, 2015, pp. 879–884.

[24] Kate Forbes-Riley and Diane 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.

[25] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Stefanos Zafeiriou, et al., "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP 2016*. IEEE, 2016, pp. 5200–5204.

[26] WQ Zheng, JS Yu, and YX Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," in *Proceedings of ACII 2015*. IEEE, 2015, pp. 827–831.

[27] Ercheng Pei, Le Yang, Dongmei Jiang, and Hichem Sahli, "Multimodal dimensional affect recognition using deep bidirectional long short-term memory recurrent neural networks," in *Proceedings of ACII 2015*. IEEE, 2015, pp. 208–214.

[28] Lang He, Dongmei Jiang, Le Yang, Ercheng Pei, Peng Wu, and Hichem Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 73–80.

[29] Shizhe Chen and Qin Jin, "Multi-modal dimensional emotion recognition using recurrent neural networks," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 49–56.

[30] Linlin Chao, Jianhua Tao, Minghao Yang, Ya Li, and Zhengqi Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 65–72.

[31] Jinkyu Lee and Ivan Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proceedings of ISCA 2015*, 2015.

[32] Yoann Baveye, Emmanuel Dellandréa, Christel Chamaret, and Liming Chen, "Deep learning vs. kernel methods: Performance for emotion prediction in videos," in *Proceedings of ACII 2015*. IEEE, 2015, pp. 77–83.

[33] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[34] Chung-Hsien Wu and Wei-Bin Liang, "Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels," *IEEE Transactions on Affective Computing*, vol. 2, no. 1, pp. 10–21, 2011.

[35] Mohammad Soleymani, Maja Pantic, and Thierry Pun, "Multimodal emotion recognition in response to videos," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 211–223, 2012.

[36] Nikolaus Bourbakis, Anna Esposito, and Despina Kavraki, "Extracting and associating meta-features for understanding people's emotional behaviour: Face and speech," *Cognitive Computation*, vol. 3, no. 3, pp. 436–448, 2011.

[37] Luciana Benotti, "Clarification potential of instructions," in *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 2009, pp. 196–205.

[38] Herbert H Clark, *Using Language*, Cambridge University Press, 1996.

[39] Myung Jong Kim, Joohong Yoo, Younggwan Kim, and Hoirin Kim, "Speech emotion classification using tree-structured sparse logistic regression," in *Proceedings of INTERSPEECH 2015*, 2015.

[40] Gary McKeown, Michel François Valstar, Roderick Cowie, and Maja Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *Proceedings of the 2010 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2010, pp. 1079–1084.

[41] Florian Eyben, Martin Wöllmer, and Björn 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.

[42] Florian Eyben, Martin Woellmer, and Bjoern Schuller, "the munich open speech and music interpretation by large space extraction toolkit," 2010.

[43] Florian Eyben, Klaus Scherer, Bjorn Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, et al., "The geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, 2015.

[44] Leimin Tian, Catherine Lai, and Johanna 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.

[45] Arman Savran, Houwei Cao, Miraj Shah, Ani Nenkova, and Ragini Verma, "Combining video, audio and lexical indicators of affect in spontaneous conversation via particle filtering," in *Proceedings of ACM-ICMI 2012*. ACM, 2012, pp. 485–492.

[46] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior research methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[47] Tom Schaul, Justin Bayer, Daan Wierstra, Yi Sun, Martin Felder, Frank Sehnke, Thomas Rückstieß, and Jürgen Schmidhuber, "PyBrain," *Journal of Machine Learning Research*, 2010.