Activity Context Representation: Techniques and Languages
AAAI Technical Report WS-12-05

Recognizing Continuous Social Engagement Level in Dyadic Conversation by Using Turn-Taking and Speech Emotion Patterns

Joey Chiao-yin Hsiao and Wan-rong Jih and Jane Yung-jen Hsu
Department of Computer Science and Information Engineering
Nation Taiwan University
Taipei,Taiwan
Email: {r99922012,wrjih,yjhsu}@csie.ntu.edu.tw

Abstract

Recognizing social interests plays an important role of aiding human-computer interaction and human collaborative works. The recognition of social interest could be of great help to determine the smoothness of the interaction, which could be an indicator for group work performance and relationship. From socio-psychological theories, social engagement is the observable form of inner social interest, and represented as patterns of turn-taking and speech emotion during a face-to-face conversation. With these two kinds of features, a multi-layer learning structure is proposed to model the continuous trend of engagement. The level of engagement is classified into "high" and "low" two levels according to human-annotated score. In the result of assessing two-level engagemet, the highest accuracy of our model can reach 79.1%.

Introduction

Social activities play a big part in human's daily life. According to Dey (2001) definition, context is any information that can be used to characterize the situation of an entity. From such viewpoint, social elements are no doubt important context when the entity is humankind. Despite the fact that computers are created to improve life of human, who is a social animal, computers are social-ignorant. In other words, computers can catch the slightest changes of human actions, but have no ability to interpret social meanings from objective actions. To better address the charateristic of users' situation, social context must be taken into consideration.

To make machines aware of social context, Pentland (2005) proposed the concept of social signal processing, and has conducted a series of research. In general scheme of social signal processing, multi-modal data streams of non-verbal behavior patterns are collected via sensors, including audio, video, and digital signals such as accelerometer and infrared sensors, The captured behavior candidates include speech, proximity, body orientation, and facial expression, etc. After features of behaviors are extracted, the patterns will be interpreted as social signals with psychological theory and mathematical techniques.

Previous works of social signal processing focused on how social signals reflect collective behaviors and group dynamics, say work performance or social relationship. For example, couples' shopping interest can be predicted from how they interact with each other, since shopping of a couple is a collaborative decision making process (Kim et al. 2009). On the other hand, Olguín, Gloor, and Pentland (2009) utilized social signals to detect nurses' personal traits and their work performance, and the result can be further used for designing group organization.

Compared to the wide application area, social signals themselves have been less discussed. Therefore, from a more fine-grained view, this paper aims to analyze one of social signal, *social engagement*, in dyadic face-to-face conversations. With microphones on smart phones, we collect humans' speech behavior, and recognize engagement level from participants' turn-taking and speech emotion. The goal is trying to assess the changing process of social engagement during a dyadic social interaction, where engagement can be estimated with a flexible time length.

Social Engagement

Engagement is defined as "*the process by which two (or more) participants establish, maintain and end their perceived connection.*" (Sidner and Dzikovska 2002). Unlike simply being involved, being engaged emphasizes more on how one actively control and participate in an interaction.

Inferring engagement level of participants during face-to-face interaction can help us better understand one's inner state of social interest. Gatica-Perez (2009) defined interest and engagement in this way: *" ... interest is used to designate people's internal state related to the degree of engagement displayed, consciously or not, during social interaction."* That is, to dig out one's internal state of social interest, we have to estimate her/his degree of engagement, where non-verbal behaviors are the observable representation of engagement.

From the cooperation of participants, Choudhury (2004) measured engagement via speech turn-taking. People's turn-taking patterns are modeled as hidden Markov processes, and being estimated how their social dynamics influence each other. With the change of one's social dynamics, engagement was thus measured. Madan (2005) had utilitized this, combined with other three signals, activity level, emphasis, and mirroring, to estimate interest, attraction, and dominance.

On the other hand, engagement is also conisdered as an presentation of personal emotion (Goodwin and Goodwin 2000). Yu, Aoki, and Woodruff (2004) analyzed engagement in coversations on telephone with this idea. They first classify speech emotion in utterance level with Support Vector Machine (*SVM*), and use the classified emotions as observation in Coupled Hidden Markov Model (*CHMM*).

Methodology

To maintain stationary property of audio signals, audio features need to be extracted within a tiny window, i.e. $1/32$ second in our experiment. However, human's social and emotion state lasts for seconds, minutes, or even longer. To recognize human's social engagement, what we need is a set of features in a bigger scale, which is capable to capture behavior change tendency. Therefore, instead of directly learning from low-level acoustic features, we propose a multi-layer learning structure to summarize them into high-level features presenting humans' behavior.

The model structure is showed in Figure 1. The three layers provide features of different scale respectively. The low-level features are directly extracted from raw audio data, and then used for generating mid-level features by using several statistical funcitons within a one second sliding window. The kinds of functions and length of window varies with the high-level features we want to extract. The high-level features are extracted with more sophisticated methods. Coupled Hidden Markov Model (*CHMM*) and K-means algorithm are used to recognize patterns in mid-level feature set, and high-level feature set will be extracted from the output of CHMM and K-means algorithm.

Low-level and Mid-level Feature Extraction

In the lowest layer, features are extracted from raw audio wave files. The audio format is 8-bit encoded wave file, sampled in 8kHz. Low-level audio features are extracted within a frame containing 250 samples, i.e. 31.25ms, where 125 overlapped samples with previous frame. As a result, within one second, there will be 64 low-level feature vectors, which contains following features: volume (log energy of signal), pitch, fundamental frequency, zero-crossing rate of signal, maximum value of non-initial auto-correlation peaks, amount of non-initial auto-correlation peaks, spectral entropy, mel frequency cepstral coefficients (MFCC).

In the mid-level, low-level features within a 1-second sliding window are crossed with several mathematical functions, e.g. mean, standard deviation, and derivation, to generate a new set of features.

Turn-taking Recognition

Turn-taking behaviors have been considered an observable presentation of one's social attitude. How one organizes her/his turn-taking structure is resulted from how she/he participates in the interaction and the relation between the interlocutor (Goodwin and Goodwin 2000). Thus we want to decode turn-taking sequences from mid-level features, and apply statistical functions on the sequences to get attributes of turn-taking.

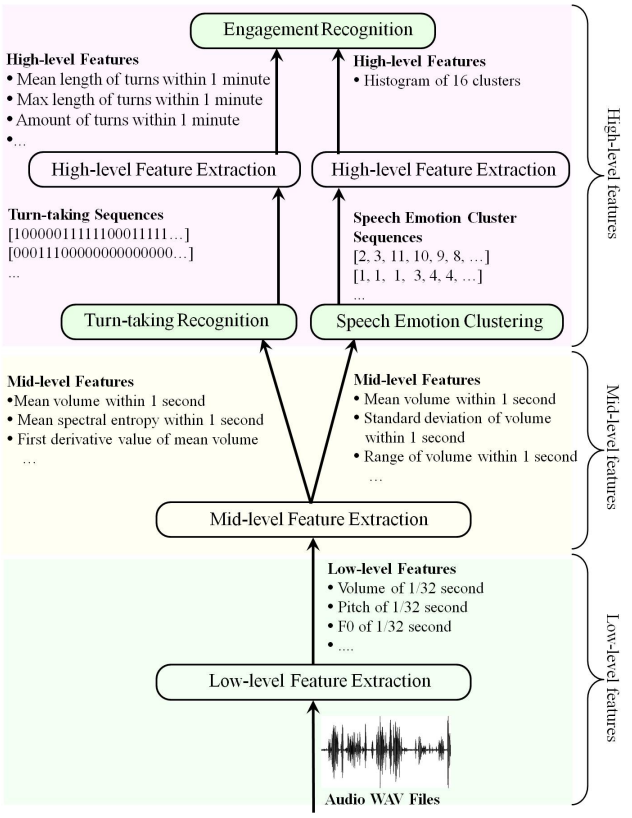

Figure 1: The hierarchical model of assessing social engagement.

In widely accepted definition, turn is a shared and limited resource, only one can have turn and speak in one time (Sacks, Schegloff, and Jefferson 1974). In every time slice, current speaker evaluates, consciously or nonconsiously, for keeping the turn or giving it up. On the other hand, a listener becomes a speaker when being assigned by current speaker or self-select (may be interruption) to take turn.

No doubt, in reality conversation goes further complicated. People use multimodal channel to commnicate and speech overlapping always happens. Kinds of models and behaviors are proposed to further clarify human conversation. To simplify the problem here, we defined that a turn is taken by one who starts speaking when no one speaks. Even her/his interacting partner starts speaking, current speaker holds turn until stopping.

CHMM is used to decode turn-taking sequences, for modeling mutual effect between paticipants. We used the features introduced by Basu (2002). For each kind of attribute, the mean value within a second is used, including: volume, maximum value of non-initial auto-correlation, number of peaks of auto-correlation, and spectral entropy, and six derivative features generated from the four features.

Acoustic Pattern Clustering

In addition to turn-taking patterns, we hope to model engagement from personal emotion. However, one challenge

of speech emotion recognition is the exhausting work of annotating. Instead of annotating emotions, we choose a simpler way, by only clustering the acoustic features which are widely used in speech emption recognition.

The low-level features used here include: volume, pitch, fundamental frequency, MFCC, and zero-crossing rate of signals. Ten mathematical functions are then applied to the low-level features within an 1-second sliding window, i.e. 32 instances, including: mean, standard deviation, variance, max, min, range (i.e. max-min), skew, kurtosis, zero-crossing rate, and mean of absolute value. Eventually, every mid-level vector contains 170 features.

K-means clustering is then used for clustering mid-level feature instances, where we assigned 16 clusters. The data will then be presented as a series of cluster numbers, e.g. [2, 3, 11, 10, 9, 8, ...].

High-level Feature Extraction

High-level Features from Turn-taking We first tried to model engagement from patterns of turn-taking. Within a 30 seconds window, high-level features of an encoded turn-taking behavior series are extracted: total length of turn, times of taking turn, mean length of turn, total length of silence, times of being silence, mean length of silence. For each participant, sign and absolute value of derivative value of the 6 features are further computed. Also, considering mutual effect between speakers, sign and absolute value of difference of the original 6 features and derivative 6 values between 2 participants are also computed. Consequently, 42 high-level features, including 18 binary and 24 numeric features are used to present the turn-taking pattern.

High-level Features from Acoustic Pattern After clustering by K-means, the mid-level features have already turned into chains containing cluster numbers. Histogram of 16 clusters within a sliding window is used as 16 features of the 30-second segment. In other words, similar histogram appears when similar speech emotion is presented, and provide the model with information to recognize level of engagement.

Engagement Recognition

Model Description Social engagement is a state shows how a participant is interested in current social interaction. Involved in a social interaction, participants' emotion and behaviors are continuous changing states, which is a result of partner's previous states and her/his previous states. To model such sequential and mutual effect of two entities, CHMM is an appropriate candidate, which models multiple Markov chains and take mutual effect into consideration.

Yu, Aoki, and Woodruff (2004) detected one's social engagement level based one utterance. Different from their work, each slice in our CHMM model is based on 30-second sliding window instead of one utterance, trying to catch behavior changes of a bigger scale than one utterance. To better model continuous behavior within a time segment, every segment overlaps 15 seconds with previous segment. Every minute could thus turn into 4 segments. By doing this, we attempt to catch more information by taking turn-taking patterns into consideration, giving more possibility in addition to only using emotion.

Annotation Instead of directly assigning engagement levels on silding windows, we assigned a score ranging from 1 to 4 scale on arbitrarily long periods, where 1 means strongly disengaged and 4 as strongly engaged.

We use voting to choose score for every second. Only when all three annotators gave different scores, the mean score would be used. Scores within a sliding window will be summed for the segment. For the 2-level engagement recognition, we assign the segment with score higher than mean of all scores as *high*, and segment with score lower than mean as *low*. By doing this way, we can have more flexibility to choose size of sliding window in future, rather than simply assigning a label onto a fixed length segment.

Experiments and Results

Data Collection

The experiment was set in a normal office environment, the participants were announced as being experiment participants, but daily activities in background were not resticted. Therefore, noise in environment would also be recorded, e.g. people's talking sound, ambient sound, or noise of daily objects. The dataset contains 11 dyadic conversations from 9 participants, collected by two iPhone 3Gs. Totally, 308 minutes of audio files are collected. The shortest length of conversation is 10 minutes, while the longest is 16 minutes, with overall average length is 14 minutes.

An one-minute sequence is separated into 4 segments, with sliding window shifts 15 seconds every time, which consequently creates total 1232 instances from 308 minutes of audio chains. The scoring task was done by three trained annotators independently. All 22 audio chains have been rated by every annotator. The annotators gave scores based on observed turn-taking behaviors and prosodic change of speakers. 94.9% of annotation received at least two agreement from three annotators.

Results of Turn-taking Recognition

To make sure the results of automatically recognized turn-taking bahavior sequences are reliable, we also conducted an experiment of turn-taking recognition.

Leave-one-sequence-out cross-validation is used to evaluate the accuracy of predicting turn-taking. The overall results of decoding the 22 chains, with 18480 seconds as total length, an average accuracy of 82.7% is reached. The best accuracy is 89.5% and the worst case is 73.8%. We believe that 82.7% is an accpetable result, thus the recognized turn-taking sequences can be futher used for extracting high-level features.

Results of Engagement Recognition

The dataset is currently small, even leaving one sequence for testing leads much information loss and results in unbalanced data. Hence instead of using leave-one-sequence-out cross validation, we choose 10-fold cross-validation to evaluate power of the learned model.

Table 1: Overall result of 2-level engagement recognition by using turn-taking patterns

	Low	High	Precision
Low	375	94	80.0%
High	255	508	66.6%
Recall	59.5%	84.4%	71.7%

Table 2: Overall result of 2-level engagement recognition by using acoustic patterns

	Low	High	Precision
Low	503	130	79.5%
High	127	472	78.8%
Recall	79.8%	78.4%	79.1%

To evaluate power of the model, an appropriate baseline is necessary for comparison. In our case, data are classified into two classes, where the amount of *low* engagement and *high* engagement instance are 630 and 602, thus 51.1% of accuracy, i.e. $630/1232$, can be used as baseline here.

Using Turn-taking Pattern Table 1 shows the results of using turn-taking patterns to recognize the engagement level, which provides 71.7% accuracy.

Using Acoustic Pattern As Table 2 shows, the experiment of 2-level engagement recognition, we can reach a 79.1% accuracy, which is the best result in our experiment.

Using Hybrid Features In addition to check the performance of each kind of feature independently, we also combined the two kinds of features to see whether they can reach a better results together. The two kinds of features mapping to the same time slice are concatenated together as one observation instance and then fed into CHMM. Table 3 is the result of using hybrid features, where the accuray is 76.7%.

Summary Compared to random guess, all results provide more than 20% accuracy, which means that the features is useful for the problem. The results of using acoustic clusters outperforms the one using turn-taking patterns and the one using hybrid features. The possible explanation is that the hitograms of acoustic clusters also contain information of turn-taking. After computed as histograms, although temporal properties of signals are faded, the histigram still provides enough information to give the model such performance. When combined with turn-taking features, the two kinds of features covered information of each other, and thus no significant improvement.

Conclusion

In this paper, we presented a method to model social engagement of participants during a face-to-face conversation under a less-controlled environment. With external sensors, people's nonverbal behavior, i.e. audio records in our experiment, can be collected without giving participants too much stress. The accuracy of 2-level engagement can reach more than 70%, which is a reliable result as first attempt. Since research in social psychology has proved that turn-taking behaviors are highly related to social engagement, we tend to conclude that more sophistacated method of feature extraction and selection may help improve the results which is the next step of this work.

Table 3: Overall result of 2-level engagement recognition by combining turn-taking patterns and audio patterns

	Low	High	Precision
Low	441	98	81.8%
High	189	504	72.7%
Recall	70.0%	83.7%	76.7%

References

Basu, S. 2002. *Conversational Scene Analysis*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Choudhury, T. K. 2004. *Sensing and modeling human networks*. Ph.D. Dissertation, Massachusetts Institute of Technology.

Dey, A. K. 2001. Understanding and using context. *Personal Ubiquitous Computing* 5.

Gatica-Perez, D. 2009. Modeling interest in face-to-face conversations from multimodal nonverbal behavior. In Thiran, J.-P.; Bourlard, H.; and Marqus, F., eds., *Multimodal Signal Processing*. Academic Press.

Goodwin, M. H., and Goodwin, C. 2000. Emotion within situated activity. In Budwig, N.; Uzgiris, I. C.; and Wertsch, J. V., eds., *Communication: An Arena of Developement*. Ablex Publishing Corporation.

Kim, T. J.; Chu, M.; Brdiczka, O.; and Begole, J. 2009. Predicting shoppers' interest from social interactions using sociometric sensors. In *Proceedings of the 27th international conference extended abstracts on Human factors in computing systems*, CHI EA '10, 4513–4518.

Madan, A. P. 2005. Thin slices of interest. Master's thesis, Massachusetts Institute of Technology.

Olguín, D. O.; Gloor, P. A.; and Pentland, A. 2009. Capturing individual and group behavior with wearable sensors. In *Proceedings of the 2009 aaai spring symposium on human behavior modeling*, SSS '09.

Pentland, A. 2005. Socially aware computation and communication. *Computer* 38:33–40.

Sacks, H.; Schegloff, E. A.; and Jefferson, G. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50(4):696–735.

Sidner, C. L., and Dzikovska, M. 2002. Human-robot interaction: Engagement between humans and robots for hosting activities. In *Proceedings of the 4th IEEE International Conference on Multimodal Interfaces*, ICMI '02, 123–128.

Yu, C.; Aoki, P. M.; and Woodruff, A. 2004. Detecting user engagement in everyday conversations. In *Proceedings of 8th International Conference on Spoken Language Processing*, ICSLP '04, 1–6.