Real-time Multimodal Affect Recognition in Laughter Episodes

A Thesis Proposal
Presented to
the Faculty of the College of Computer Studies
De La Salle University Manila

In Partial Fulfillment
of the Requirements for the Degree of
Master of Science in Computer Science

by

SANTOS, Jose Miguel

Dr. Merlin SUAREZ
Adviser

April 10, 2012

Abstract

Emotion recognition has been a widely studied subject in the literature, as knowing the emotions of other people can help a person adjust and respond appropriately. The recognition of emotion in laughter is particularly important as laughter can identify non-basic affective states such as distress, anxiety, and boredom. Existing laughter recognition systems are unable to detect laughter in real-time, however. This research proposes a system that can recognize and detect laughter in real-time, using body movements as additional features to laughter.

Keywords: Tracking, Signal processing, Computer vision, Model classification, Laughter, Affect recognition, Laughter, Gesture, Multimodal

Table of Contents

List of Figures

List of Tables

1 Research Description

This chapter introduces the research. The chapter starts with an overview of the current state of affect modeling, gesture recognition, and laughter recognition. The chapter then presents the research problem, research objectives and the scope of the research. The significance of the study is discussed as well.

1.1 Overview of the Current State of Technology

Emotion recognition has been a widely studied subject for a variety of reasons. Knowing the emotions of other people can help a person adjust and respond appropriately (Goleman, 1995). Changes in emotion is also a fundamental component of human-to-human communication (Zeng et al., 2009). Moreover, research has shown a clear link between motivations of people and the emotions they feel (Rolls, 2007). Thus, recognizing emotions is key to both understanding the motivations of people and responding appropriately to those motivations. Emotion recognition can also help in determining which emotions are universal, and which emotions are just cultural (Elfenbein and Ambady, 2002). Studies of emotion recognition have even extended to autistic spectrum disorders (Loveland et al., 1997) and schizophrenia (Li et al., 2010).

Affective computing is computing that relates to, arises from, or influences emotions (Picard, 1995). Plenty of research has been done on affective computing, with varying methodologies. Researches in the field can take the single modality approach, using a single input source such as spoken words (Huber et al., 2000; Fernandez, 2004; Zeng et al., 2004), facial expressions (Sarrafzadeh et al., 2003; Hu et al., 2008; Zeng et al., 2004), body movements or laughter. Other researches opt to use a multi-modal approach, where more than one input source is used. For example, some researches use a combination of facial expressions, body gestures, conversational cues or computer activity (Gunes and Piccardi, 2007; Kapoor and Picard, 2005; D'Mello and Graesser, 2010; Zeng et al., 2004), or a combination of different physiological signals (Nasoz et al., 2004; Wang et al., 2004; Mandryk and Atkins, 2007; Zhai and Barreto, 2006).

Gestures in particular are often used as an input source for affect modeling (Gunes and Piccardi, 2007; De Silva et al., 2006). One particular research hypothesized that certain body movements help a person cope with an experienced emotion (Ekman and Friesen, 1974). In other words, certain body movements could actually indicate the affect of a subject.

A lot of the research done in gesture-based affect modeling take a marker-based approach such as wires (De Silva et al., 2006) and reflective tapes (Kapur et al., 2005; Atkinson et al., 2004) to track body motion. This approach is not ideal as it is obtrusive and distracting, inhibiting spontaneity due to discomfort on the side of the subject. Other research in the field use the markerless approach, where no sensors are attached to the body of the subject. These research rely only on video recordings and template matching (Rosenhahn et al., 2008; Yoo and Nixon, 2011), or data gathered from specialized cameras like stereo cameras (Lin et al., 2010) or the Microsoft Kinect (Kristensson et al., 2012).

Audial information like laughter and speech are also used as input sources for affect modeling. This is because studies show that pitch and energy are useful in recognizing the affect state of the subject (Zeng et al., 2009), thus making them valuable modalities for affect modeling.

Laughter is particularly important as it is considered one of the most noteworthy paralinguistic sounds (Escalera et al., 2009). Non-linguistic vocalizations like laughter are also significant because they can identify non-basic affective states such as distress, anxiety, and boredom (Zeng et al., 2009). Laughter can be categorized into different types, such as joy, tickling, taunting and schadenfraude, which is the pleasure in another's misfortune (Szameitat et al., 2009).

Data corpuses of researches in laughter fall into two categories: acted laughter (Urbain et al., 2010), and spontaneous laughter (Devillers and Vidrascu, 2007; Pantic and Petridis, 2008). For spontaneous laughter, instances in the corpus are taken from conversations between people, either from formal settings like work (Devillers and Vidrascu, 2007; Pantic and Petridis, 2008) or casual conversations (Bantiling et al., 2010). In one research, one of the subjects is given a strict strategy during the interaction, and the conversations are classified by the subjects into a discrete set of topics. As the conversations take place in the context of a call, the conversations were easy to record. The instances were hand-transcribed and segmented, and given an emotion label, which was either positive or negative (Devillers and Vidrascu, 2007).

Only a small number of studies concerning automatic detection of paralinguistic vocalizations, like laughter, have been available (Zeng et al., 2009). Also, most studies made for human affect analysis were done using a single-modal approach (Zeng et al., 2009). Not all emotions can be accurately determined through gestures alone (Bustos et al., 2011). Thus, further research is still needed using gestures in affect modeling, particularly using a multimodal approach. Furthermore, existing laughter recognition systems are unable to detect laughter in real-time Alonzo et al. (2010); Galvan et al. (2011).

1.2 Research Objectives

1.2.1 General Objective

In light of the findings presented above, this research aims to answer the question:

HOW CAN THE EMOTION OF LAUGHTER BE RECOGNIZED IN REAL-TIME?

1.2.2 Specific Objectives

Specifically, the following sub-problems must be answered:

1. **How can a corpus of spontaneous gestures and laughter be built?** The researcher needs to identify the gestures spontaneously performed by subjects while they are in the middle of a laughter episode. The researcher must also identify how to handle the data gathering for this research, keeping in mind that the laughter episodes must be spontaneous and genuine, and that visual information must be gathered at the same time.

2. **How can laughter be recognized in real-time?** To be able to determine the affect of laughter in real-time, the researcher must first be able to construct a system that can detect laughter in real-time. The system should be able to receive as input a stream of a conversation, and mark the times when laughter episodes start, and when laughter episodes end. The system should not be limited to manually segmented laughter episodes.

3. **What features should be used to build the emotion classifier?** The features of laughter relevant to emotion recognition is still unknown, and thus the researcher must first be able to determine said features. These features will be decided from the results of experiments performed throughout the research, with the best-performing features relative to classification accuracy chosen as the relevant features.

1.3 Scope and Limitations of the Research

The research will only deal with spontaneous laughter, or laughter that is not acted. More specifically, the research will be concerned with laughter that naturally occurs during conversation. Other research in the field, like Devillers and

Vidrascu (2007), also use spontaneous laughter. In addition, only voiced laughter will be considered, due to limitations that a real-time system would entail. This is because the system would need cues to know when laughter starts and laughter ends, which unvoiced laughter lack.

Likewise, only spontaneous gestures, or non-acted gestures, are considered. In addition, the gestures will only be limited to upper body movements, or gestures related to the head, shoulders and hands. The system being built will only track the head position, the shoulder positions, and the hand positions.

While the system is real-time, the system will not return immediate results but rather results that are processed shortly after the significant events happen. This is due to the nature of laughter, which has to be complete before it can be processed.

For the data corpus, subjects will strictly be students of De La Salle University who are studying under the College of Computer Science. As the research will focus on spontaneous laughter, the researcher will use Devillers and Vidrascu (2007) as basis. Specifically, the corpus will compose of laughter episodes taken from video files of subjects engaged in conversation. For data gathering, each session will require two subjects, who are acquaintances and have known each other for at least 6 months, and are conversing with each other using video chatting software. To ensure that either subject will engage in laughter during the conversion, the topic of the conversation will be of a humorous nature, chosen by the subjects from a list provided by the researcher. The list contains topics that are generally considered humorous. Specifically, these topics are comedy television shows, inside jokes, and funny habits by acquaintances.

In addition, for both subjects the upper body and face must be visible at all times. This is so the upper-body gestures required by the research will be captured. The lengths of each session are variable, but must not exceed 1 hour. This is so that conversations have enough time to play out to their natural end. The laughter episodes will be manually segmented from the session videos, and an event will be considered a laughter episode based on the input of the subject. That is, the subject will be shown the videos post-recording and are asked to annotate events that he or she considers as a laughter episode. This is as the subjects are the most qualified to say what emotions they were feeling. Laughter episodes have variable lengths as well.

Each laughter episode will also be labelled by the subjects post-recording, no more than a day after the video was first recorded. This is so the subjects would still remember the emotions that they were feeling. For the emotion labels, the abstract dimensions method used by Galvan et al. (2011), which use valence and arousal values to represent emotion, will be used.

1.4 Significance of the Research

This research will produce a system that can detect laughter in real-time, which is important in systems like embodied conversation agents (ECA) that interact with users in real-time. These systems are limited by the information they can process. If they can only use gestures for emotion recognition, for example, the quality of the classification model will not be as effective as a multi-modal one. Thus, being able to recognize laughter in real-time will greatly improve emotion recognition and even the breadth of emotions under the scope of these systems.

The affect of laughter on its own is not well studied in the literature, often being combined with other vocal input like speech. This is not very ideal, as laughter is quite unique from other vocalities. Laughter that sounds positive and happy can actually mean the very opposite, or should be taken in a sarcastic tone. Not all laughter is voiced, as well. Thus, there is significance in studying the affect of laughter alone.

2 Review of Related Literature

This chapter discusses the related works and systems in recognizing affect through laughter and gestures. This chapter is divided into four sections: Section 2.1 Laughter Segmentation and Detection Systems; Section 2.2 Affective Modeling Using Laughter; Section 2.3 Affective Modeling Using Gestures; and Section 2.4 Multimodal Affect Modeling.

2.1 Laughter Segmentation and Detection Systems

A lot of research has been done on laughter detection systems. The researches mostly fall into two categories: laughter-versus-speech discrimination systems, where laughter episodes are pre-segmented and the aim is to classify these segmented episodes (Lockerd and Mueller, 2002; Schuller et al., 2008; Truong and van Leeuwen, 2007b); and laughter segmentation systems, where the aim is to segment the laughter episodes from the audio stream into laughter and non-laughter episodes (Kennedy and Ellis, 2004; Knox et al., 2008; Laskowski and Schultz, 2008).

The researches also take different approaches for the classifier model used. Some researches use static models, which classify by frame, like neural networks (Petridis and Pantic, 2008; Knox et al., 2008) and support vector machines (SVM) (Kennedy and Ellis, 2004; Truong and van Leeuwen, 2007a), while other researches use dynamic models that classify by sequence, like Hidden Markov Models (HMM) (Lockerd and Mueller, 2002; Campbell et al., 2005; Laskowski and Schultz, 2008) and Gaussian Mixture Models (GMM) (Truong and van Leeuwen, 2007b). Schuller et al. (2008) discovered that SVMs are comparable in performance and accuracy to HMMs for classifying nonlinguistic vocalizations. In addition, according to Petridis et al. (2009), neural networks in general are comparable to HMMs when classifying laughter versus speech.

For testing the classifier models, some researches use some form of cross-validation Petridis and Pantic (2008); Schuller et al. (2008), or use a different set for testing than the one used for training (Knox et al., 2008; Truong and van Leeuwen, 2007b; Laskowski and Schultz, 2008).

Laskowski and Schultz (2008) used silence, speech and laughter as the labels for the instances in the corpus.

Many of the researches used pre-existing data corpus for training and testing. These include the ICSI dataset, which is composed of natural meeting recordings where the subjects are wearing head-worn microphones in addition to various desktop microphones placed around the table. Each meeting had six participants, and only audio information was recorded (Janine et al., 2003). Researches that used the ICSI dataset include Knox et al. (2008), Kennedy and Ellis (2004), and Truong and van Leeuwen (2007b).

Another common dataset is the AMI dataset by McCowan et al. (2005), which is composed of recordings of real meetings and scenario-driven meetings. In addition, the corpus contains visual and audio information. A screen shot of an instance in the AMI corpus can be seen in Figure 2.1. All the participants in the dataset are non-native English speakers. Researches that used the AMI dataset include Petridis and Pantic (2008), Petridis and Pantic (2011) and Reuderink et al. (2008).s

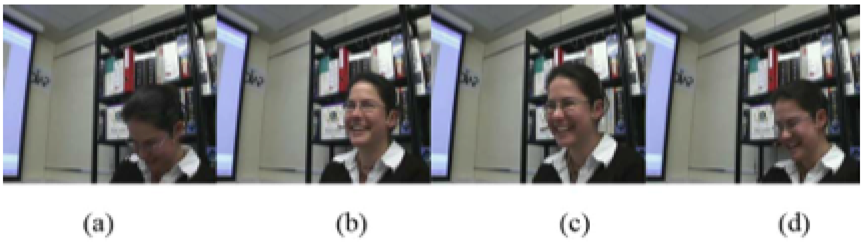

Figure 2.1: Screenshot of a voiced laughter instance in the AMI dataset (Petridis and Pantic, 2011)

Another dataset is the SAL dataset, which is made up of recordings of conversations between a subject and the SAL agent. The laughter episodes were already pre-segmented, and audio and visual information were both recorded. An screen shot of an instance in the SAL corpus can be seen in Figure 2.2. Petridis and Pantic (2011) used the SAL dataset for one of their experiments.

Other researches, like Bantiling et al. (2010), built their own corpus. They recorded meetings in a casual context, and the laughter episodes were pre-segmented.

A lot of the research in laughter detection focus on cepstral features and prosodic features (Truong and van Leeuwen, 2007a; Schuller et al., 2008; Petridis and Pantic, 2008; Kennedy and Ellis, 2004; Laskowski and Schultz, 2008; Knox et al., 2008). Cepstral features include mel-frequency cepstral coefficients (MFCC) and perceptral linear predictive (PLP). Most researches use 12 or 13 coefficients (Petridis and Pantic, 2011), however according to Yin et al. (2006), using six or seven coefficients produce similar or better results to 12 or 13 coefficients. The prosodic features commonly used in laughter detection research are pitch and energy (Zeng et al., 2009). According to Bachorowski et al. (2001), pitch is higher

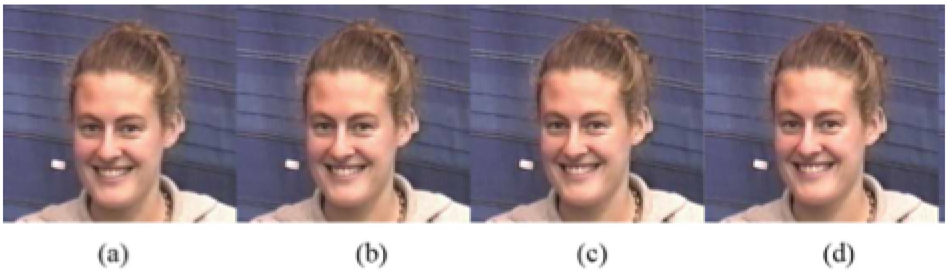

Figure 2.2: Screenshot of a voiced laughter instance in the SAL dataset (Petridis and Pantic, 2011)

during laughter than during speech, and thus is very helpful in distinguishing between the two. Some researches combine the cepstral and prosodic features at the feature level (Yin et al., 2006) or the decision level (Graciarena et al., 2006). Furthermore, some research combine these cepstral and prosodic features with other modalities, such as facial expressions, in a multi-modal approach (Petridis and Pantic, 2008, 2011).

Petridis and Pantic (2011) did several experiments, using different combinations of features. One set of features used was the facial points of the subject, an example of which can be seen in Figure 2.3. As can be seen in Figure 2.3, they only track a limited number of points, specifically the points around the eyes, eyebrows, nose, mouth and the jaw. They tried the single-modal approach, and using just facial features achieved a classification rate of 83.9%. Using the cepstral and prosodic features, they achieved a classification rate of 92.3%. They also tried the bi-modal approach, and got a classification rate of 94.4% by combining facial features and MFCC features.

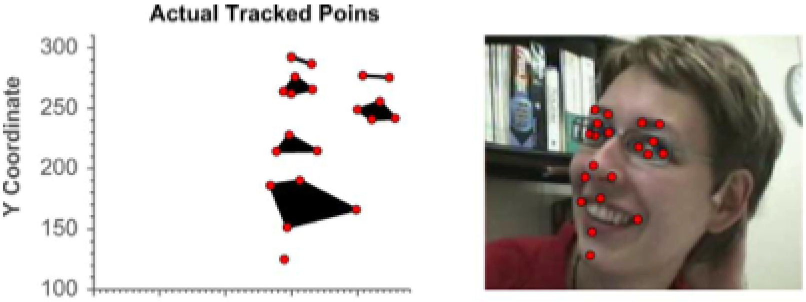

Figure 2.3: Actual tracked points from the AMI dataset (Petridis and Pantic, 2011)

Knox et al. (2008) created a laughter segmentation system, and used neural networks as their classification model. Prosodic and cepstral features were used, and the system achieved an equal error rate (EER) of 5.4%. Lockerd and Mueller (2002) achieved an 88% classification rate on their laughter-versus-speech (LVS) discrimination system using HMMs and spectral features. Petridis and Pantic (2008) also made a LVS discrimination system, but used neural networks as the classification model, and was able to achieve a recall of 86.9% and a precision of 76.7% using facial points distance and PLP. Finally, Truong and van Leeuwen (2007a) used SVM and GMM for their LVS discrimination system, achieving an EER of 2.8% for their user-specific model.

2.2 Affect Modeling Using Laughter

Not a lot of literature exists on modeling affect specifically with laughter. Instead, most research focus more on modeling affect using voice or speech, which can include laughter as part of the scope (Schuller, 2011; Huang and Changxue, 2006; Banziger et al., 2009).

Schuller (2011) discussed the principles on how to analyze and handle audio information for affect recognition. Different methodologies could be used to separate the main audio stream into "chunks": there is the "window" approach where the stream is broken down into set intervals, and the "turn" approach where the stream is broken down based on speech onset until offset by checking if the energy level exceeds a certain threshold. For feature extraction, the prosodic and cepstral features can be used. For classification, linear discriminant classifiers (LDC), k-Nearest Neighbor (kNN), SVM, HMM, neural networks can all be used, and can even be combined using algorithms like MultiBoosting and Stacking.

Huang and Changxue (2006) built an affect detection system that runs in real-time, and ignores linguistic and semantic information in the audio, focusing only on the acoustic level. They used continuous density HMM as the classification model as it is sequential in nature, and they believe capturing the fluctuation of features is important in detecting emotion. They also used pitch, energy, EP and energy slope as features, because they are segmental features and are more relevant in a sequential model. GMMs were used to capture the range, mean, median and variability of the pitch and energy, and the results were fed to each state in the HMM. An HMM classifier was built for each emotion, and the model with the highest probability will be the assigned emotion to the input audio. They were able to achieve an accuracy of 98% when discriminating neutral emotions from angry emotions, 69% when discriminating neutral emotions from sad emotions, and 82% when discriminating angry emotions from happy emotions. Table 2.1

shows the results versus another research. The features used in each experiment is indicated in the number inside the parentheses next to the experiments names. The system of Huang and Changxue (2006), represented by the columns labelled EP, ZEP and ZEPS, were able to outperform the system build by HP-Labs, despite the fewer number of features. Most notably, they more than doubled the accuracy when it comes to predicting all emotions, and improved from 50% to 69% when discriminating sad versus angry, which is significant because the accuracy can no longer be attributed to just chance.

Table 2.1: Emotion detection accuracy of the system (Huang and Changxue, 2006)

	Neutral vs. Angry	Neutral vs. Sad	Angry vs. Happy	All Emotions
HP-Labs(37)	94.5%	50%	?	8.7%
EP (2)	95%	54%	74%	13%
ZEP (3)	98%	62%	77%	15%
ZEPS (4)	98%	69%	82%	18%

Devillers and Vidrascu (2007) used a natural corpus made up of spontaneous dialog for their system. They used discrete classes of laughter, which were positive, negative and ambiguous, for the labels and used the mean and standard deviation of F0 statistics, the percentage of unvoiced frames, and the energy and duration of each laughter episode as the features. The results of the study were that unvoiced laughter were more common in negative emotion laughter than in positive emotion laughter.

Banziger et al. (2009) stated that using a small number of emotions for labels, specifically less than 10, is bad as the model is more susceptible to discrimination and guessing. They suggested anxiety, panic, fear, happiness, elation, cold anger, hot anger, sadness, despair, disgust, and contempt; this is because each of the five major emotion families is represented by two variants.

2.3 Affect Modeling Using Gestures

Few researches in the field of affect recognition focus on using gestures as the only modality. Some of the researches can be quite helpful, however, in determining how gestures and their features should be handled by an affect recognition system.

A study by D'Mello and Graesser (2009) aimed to detect a subject's affect using the subject's gross body language while interacting with the AutoTutor Intelligent Tutoring System (ITS). An automated body pressure measurement system was used to capture the pressure that the subject was exerting on the chair, in effect measuring the posture and body motions that the subject was doing. The corpus was created by capturing interactions of different subjects with AutoTutor, and annotated with an emotion by different people: the subject, a peer, and two trained judges. Two different sets of features were collected, one being the measurement of the actual pressure exerted, including the magnitude and direction of the pressure. The other set focused on the spatial and temporal characteristics of the pressure exerted. Five different datasets were created, with the first four datasets being the integration of the features collected with the temporal affective characteristics given by the subject, peer, and trained judges, and the last dataset being the integration with both judges. The results of the experiments can be seen in Table 2.2. They were able to attain accuracies well above chance, which is 50%, suggesting that the classification model works fairly well.

Table 2.2: Affect detection accuracies versus neutral (D'Mello and Graesser, 2009)

Emotion	**Accuracy**
Boredom	73%
Confusion	72%
Delight	70%
Flow	83%
Frustration	74%

2.4 Multimodal Affect Modeling

Plenty of research has been done on multimodal affect modeling, often combining video information and audio information such as facial features and speech in order to achieve a better classification rate (Cueva et al., 2011; Metallinou et al., 2010).

In a study by Alonzo et al. (2010), a multimodal system using facial expressions and vocal features was built to classify the affect of laughter. The corpus was composed of acted laughter. For the vocal features, the pitch, intensity, formants, pitch contour points and MFCC were extracted from the instances, along with the mean, minimum value, maximum value and standard deviation of the the pitch and intensity. For the facial expression features, the facial distances between 64 facial points were extracted. MLP, kNN and SVM were used as the classification algorithms, and decision-level fusion was used to combine the modalities. The results of their study show that prosodic features are more indicative of the affect, and that pitch contour should be excluded from the prosodic features to get better results.

Cueva et al. (2011) built a multimodal system by fusing facial and vocal features in the decision level. They used eMotion to interpret the emotion of the facial features, and Emo-Voice to interpret the vocal features, and emoCrawler to interpret the semantic features. They used the eNTERFACE'05 Audio-Visual Emotion Detection which contained acted instances, and separated the corpus into three different sets, one set for training with Emo-Voice, another set for training using the fusion of all modalities, and the last set for testing. The training set was purposely left noisy, and neural networks were used for fusion as they can gauge the importance of certain input, which is important in determining which modalities contribute more to the emotion. The results of the experiments can be seen in Table 2.3. Fusing the facial features with the vocal features and semantics greatly improved the accuracy in terms of predicting happiness and fear. On the other hand, it actually diminished the accuracy in predicting sadness, relative to using voice only.

Table 2.3: Rate of detection for each method (Cueva et al., 2011)

Emotion	Voice	Face	FFBP(F+V+Sem)	PNN(F+V+Sem)
Happiness	20%	0%	60%	60%
Anger	100%	0%	100%	100%
Fear	40%	20%	80%	60%
Sadness	100%	60%	60%	60%
Average Rate	**98%**	**69%**	**82%**	**18%**

Metallinou et al. (2010) also built a multimodal system using facial and vocal features, and like Cueva et al. (2011) employed decision-level fusion. A classifier was built for each modality, and they applied a Bayesian framework to combine the different classifiers, with in-domain information to enrich the fusion process. The IEMOCAP corpus was used, which is composed of scripted sessions between actors. Facial features were composed of coordinates of markers connected to the face, which were reduced by feature selection algorithms to avoid correlation between features. Examples of the markers can be seen in Figure 2.4. Vocal features were composed of 13 Mel Filterbank Coefficients (MFB), as they have shown to perform better on emotion recognition tasks (Busso et al., 2007), in addition to pitch and energy values.

Using the decisions of each modality, a histogram containing the amount of time that each instance is classified to a certain emotion is created, and is used to approximate the probability that an instance belongs to a particular emotion. The results of the experiment using 10-fold leave-one-speaker-out cross validation

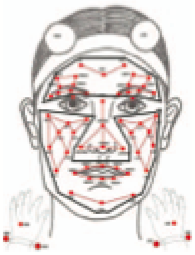

Figure 2.4: Face and head marker positions (Metallinou et al., 2010)

can be seen in Table 2.4. The higher accuracies when separating the upper part face from the lower part face suggest that the different parts of the face have different influences on affect. The differences between the experiments are quite small, however, and thus no conclusion can really be reached.

Table 2.4: Percentages per utterance (Metallinou et al., 2010)

Bayes Fusion	**Total Unweighted Accuracy**
Face + Voice	60.57% ± 4.26%
UpperFace + LowerFace + Voice	61.15% ± 3.62%
Face + Voice + Head	61.79% ± 3.96%
UpperFace + LowerFace + Voice + Head	62.42% ± 3.16%

3 Research Methodology

This chapter shall enumerate the phases of the study and the specific tasks involved in each phase. The phases occur in a sequential manner, though some phases remain constant throughout like Documentation, while other phases may be revisited as new limitations and breakthroughs are discovered, or just to refine the existing systems.

3.1 Concept Formulation and Review of Related Literature

In this phase, the details of the research, including the scope, methodology, and objectives will be determined. Related literature will be studied to understand what problems need to be solved, and the limitations and approaches of other similar studies. It is also in this phase that possible solutions to the discovered problems will be researched, and even partly experimented on. Systems and tools that will be used during the research will also be learned during this phase.

3.2 Prototype Building

In this phase, a prototype system will be built to extract vocal features like cepstral and prosodic features in real-time and test the viability of classifying laughter in real-time. Various approaches will be tested and experimented to see which is the most viable way of doing the system in real-time. A stand-in classification model will be used in place of the actual classification model which has not been built yet.

3.3 Data Collection and Corpus Building

During this phase, data for training and testing will be collected to fit the established needs of the system and classification model. Volunteers will be gathered from students in the College of Computer Studies at De La Salle University, and be subject to the data collection steps outlined in the scope and limitations. All the features that can be extracted from the recorded sessions shall be extracted, and redundant features will simply be excluded in the 3.4 section. This is so that should a feature turn out to be needed, data will not have to be recollected. This phase will reoccur constantly throughout the research, to continue to grow the data corpus in case more sessions will be needed and make the corpus more diverse and representative.

3.4 Training and Evaluation

The classification model of the system will be built during this phase. The building of the feature set, adjustment of features excluded, adjustment of labels to use and adjustment of algorithms to be used will all take place in this phase. The results of each model built will also be evaluated during this phase, and evaluation includes analyzing which classifiers do best, which set of features are most important, and which labels should be used to gain a better accuracy. This phase will also occur more than once, as more data and better suited features and algorithms are discovered.

3.5 Documentation

The documentation phase shall be done throughout the research. Taking down notes, writing about the related literature, building technical documents and theoretical frameworks, creating test cases and writing the thesis document all fall under the documentation phase.

3.6 Calendar of Activities

Table 3.1 shows the schedule of activities for the year 2012. Each bullet represents approximately one week worth of activity.

Table 3.1: Timetable of Activities

Activities (2012)	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Concept Formulation and Review of Related Literature	●●●●	●●●●	●●●●	●●●								
Prototype Building				●	●●●●							
Data Collection and Corpus Building						●●●●	●●●●		●●	●●●●		
Training and Evaluation								●●●●	●●		●●●●	●●
Documentation	●●●●	●●●●	●●●●	●●●●	●●●●	●●●●	●●●●	●●●●	●●●●	●●●●	●●●●	●●

References

Alonzo, J. Campita, J., Lucila, S., and Miranda, M. (2010). Discovering emotion in filipino laughter using multimodal approaches. Undergraduate Thesis, De La Salle University Manila.

Atkinson, A., Dittrich, W., Gemmell, A., and Young, A. (2004). Emotion perception from dynamic and static body expressions in point-light and full-light displays. *Perception*, 33:717–746.

Bachorowski, J. A., Smoski, M. J., and Owren, M. J. (2001). The acoustic features of human laughter. *J. Acoust. Soc. Amer.*, 110:1581–1697.

Bantiling, H. P., Gadi, S. R., Lee, J. C., and Yang, J. V. (2010). Automatic video segmentation tool for laughter detection based on audio features. Undergraduate Thesis, De La Salle University Manila.

Banziger, T., Grandjean, D., and Scherer, K. (2009). Emotion recognition from expressions in face, voice, and body: The multimodal emotion recognition test (mert). *Emotion*, 9:691–704.

Busso, C., Lee, S., and Narayanan, S. S. (2007). Using neutral speech models for emotional speech analysis. In *Interspeech 2007*.

Bustos, D. M., Chua, G. L., Cruz, R. T., and Santos, J. M. (2011). Markerless gesture recognition in the context of affect modeling for intelligent tutoring systems. Undergraduate Thesis, De La Salle University Manila.

Campbell, N., Kashioka, H., and Ohara, R. (2005). No laughing matter. *Proc. Eur. Conf. Speech Communication and Technology*, pages 465–468.

Cueva, D., Goncalves, R., Cozman, F., and Pereira-Barretto, M. (2011). Crawling to improve multimodal emotion detection. *Lecture Notes in Artificial Intelligence*, 7095.

De Silva, P., Madurapperuma, A., Marasinghe, A., and Osano, M. (2006). A multi-agent based interactive system towards child amp; no.146; emotion performances quantified through affective body gestures. In *Pattern recognition.*

Devillers, L. and Vidrascu, L. (2007). Positive and negative emotional states behind the laughs in spontaneous spoken dialogs. In *Interdisciplinary Workshop on The Phonetics of Laughter*, pages 37–40.

D'Mello, S. and Graesser, A. (2009). Automatic detection of learner's affect from gross body language. *Applied Artificial Intelligence*, 23:123–150.

D'Mello, S. and Graesser, A. (2010). Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. *User Modeling and User-Adapted Interaction*, 20:147–187.

Ekman, P. and Friesen, W. (1974). Detecting deception from the body and face. In *Journal of Personality and social psychology*, pages 288–298.

Elfenbein, H. A. and Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. In *Psychological Bulletin*.

Escalera, S., Puertas, E., Pujol, O., and Radeva, P. (2009). Multi-modal laughter recognition in video conversations. In *Computer Vision and Pattern Recognition Workshops*.

Fernandez, R. (2004). A computational model for the automatic recognition of affect in speech.

Galvan, C., Manangan, D., Sanchez, M., and Wong, J. (2011). Audiovisual affect recognition in spontaneous filipino laughter. (Undergraduate Thesis).

Goleman, D. (1995). *Emotional intelligence*. Bantam Books.

Graciarena, M., Shriberg, E., Stolcke, A., Enos, F., Hirschberg, J., and Kajarekar, S. (2006). Combining prosodic lexical and cepstral systems for deceptive speech detection. *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1:1033–1036.

Gunes, H. and Piccardi, M. (2007). Fusing face and body gesture for machine recognition of emotions.

Hu, Y., Zeng, Z., Yin, L., Wei, X., Tu, J., and Huang, T. (2008). A study of non-frontal-view facial expressions recognition. In *Pattern recognition*.

Huang, R. and Changxue, M. (2006). Toward a speaker-independent real-time affect detection system. *Pattern Recognition*, pages 1204–1207.

Huber, R., Batliner, A., Buckow, K., Noth, E., Warnke, V., and Niemann, H. (2000). Recognition of emotion in a realistic dialogue scenario. In *Spoken language processing*.

Janine, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., and Wooters, C. (2003). The icsi meeting corpus. *Proc. IE*, 1:364–367.

Kapoor, A. and Picard, R. W. (2005). Multimodal affect recognition in learning environments. *MULTIMEDIA '05*, pages 677–682.

Kapur, A., Virji-Babul, N., Tzanetakis, G., and Driessen, P. (2005). Gesture-based affective computing on motion capture data. In *Proc. int. conf. affective computing and intelligent interaction*, pages 1–7.

Kennedy, L. and Ellis, D. (2004). Laughter detection in meetings. In *Proc. NIST Meeting Recognition Workshop*.

Knox, M., Morgan, N., and Mirghafori, N. (2008). Getting the last laugh: Automatic laughter segmentation in meetings. *Proc. INTERSPEECH*, pages 797–800.

Kristensson, P. O., Nicholson, T., and Quigley, A. (2012). Continuous recognition of one-handed and two-handed gestures using 3d full-body motion tracking sensors.

Laskowski, K. and Schultz, T. (2008). Detection of laughter-in-interaction in multichannel close-talk microphone recordings of meetings. *Lec*, 5237:149–160.

Li, H., Chan, R., McAlonan, G., and Gon, Q.-y. (2010). Facial emotion processing in schizophrenia: a meta-analysis of functional neuroimaging data. *Schizophrenia Bulletin*, 36:1029–1039.

Lin, S.-Y., Lai, Y.-C., Chan, L.-W., and Hung, Y.-P. (2010). Real-time 3d model-based gesture tracking for multimedia control. In *International Conference on Pattern Recognition*.

Lockerd, A. and Mueller, F. (2002). Lafcam: Leveraging affective feedback camcorder. *CHI 2002*, pages 574–575.

Loveland, K., Tunali-Kotoski, B., Chen, Y. R., Ortegon, J., Pearson, D., Brelsford, K., and Gibbs, M. C. (1997). Emotion recognition in autism: Verbal and nonverbal information. *Development and Psychopathology*, 9:579–593.

Mandryk, R. and Atkins, M. S. (2007). A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *Int. J. Human-Computer Studies*, 65:329–347.

McCowan, I., Carletta, J., Kraaji, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J. ans Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., and Wellner, P. (2005). The ami meeting corpus. *Proc. Int. Conf. Methods and Techniques in Behavioral Research*, pages 137–140.

Metallinou, A., Lee, S., and Narayanan, S. (2010). Decision level combination of multiple modalities for recognition and analysis of emotional expression. *ICASSP 2010*, pages 2462–2465.

Nasoz, F., Lisetti, C., Alvarez, K., and Finkelstein, N. (2004). Emotion recognition from physiological signals for user modeling of affect. *Cognition, Technology & Work*, pages 4–14.

Pantic, M. and Petridis, S. (2008). Audiovisual discrimination between laughter and speech. In *Acoustics, Speech and Signal Processing*, pages 5117–5120.

Petridis, S., Gunes, H., Kaltwang, S., and Pantic, M. (2009). Static vs. dynamic modeling of human nonverbal behavior from multiple cues and modalities. *Proc. ICMI*, pages 23–30.

Petridis, S. and Pantic, M. (2008). Audiovisual discrimination between laughter and speech. *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, pages 5117–5120.

Petridis, S. and Pantic, M. (2011). Audiovisual discrimination between speech and laughter. *IEEE Transactions on Multimedia*, 13.

Picard, R. (1995). Affective computing. Technical report, M.I.T Media Laboratory Perceptual Computing Section.

Reuderink, B., Poel, M., Truong, K., Poppe, R., and Pantic, M. (2008). Decision-level fusion for audio-visual laughter detection,. *Lecture Notes in Computer Science*, 5237:137–148.

Rolls, E. T. (2007). *Emotion explained*. Oxford University Press.

Rosenhahn, B., Schmaltz, C., Brox, T., Weickert, J., Cremers, D., and Seidel, H.-P. (2008). Markerless motion capture of man-machine interaction. pages 1–8.

Sarrafzadeh, A., Hosseini, H., Fan, C., and Overmyer, S. (2003). Facial expression analysis for estimating learner's emotional state in intelligent tutoring systems. In *Advanced learning technologies*.

Schuller, B. (2011). Voice and speech analysis in search of states and traits. *Computer Analysis of Human Behavior*, pages 227–253.

Schuller, B., Eyben, F., and Rigoll, G. (2008). Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech. *Lecture Notes in Computer Science*, 5078:99–110.

Szameitat, D. P., Alter, K., Szameitat, A. J., Darwin, C. J., Wildgruber, D., Dietrich, S., and Sterr, A. (2009). Differentiation of emotions in laughter at the behavioral level. *Emotion 2009*, 9:397–405.

Truong, K. P. and van Leeuwen, D. A. (2007a). Automatic discrimination between laughter and speech. *Speech commun.*, 49:144–158.

Truong, K. P. and van Leeuwen, D. A. (2007b). Evaluating automatic laughter segmentation in meetings using acoustic and acoustic-phonetic features. In *Proc. Workshop Phonetics of Laughter*.

Urbain, J., Bevacque, E., Dutoit, T., Moinet, A., Niewiadomski, R., Pelachaud, C., Picart, B., TIlmanne, J., and Wagner, J. (2010). The avlaughtercycle database.

Wang, H., Predinger, H., and Igarashi, T. (2004). Communicating emotions in online chat using physiological sensors and animated text.

Yin, B., Ambikairajah, E., and Chen, F. (2006). Combining cepstral and prosodic features in language identification. *Proc. Int. Conf. Pattern Recognition*, 4:254–257.

Yoo, J.-H. and Nixon, M. (2011). Automated markerless analysis of human gait motion for recognition and classification.

Zeng, Z., Pantic, M., Roisman, G., and Huang, T. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31.

Zeng, Z., Tu, J., Liu, M., Zhang, T., Rizzolo, N., Zhang, Z., Huang, T., Roth, D., and Levinson, S. (2004). Bimodal hci-related affect recognition. *ICMI '04*, pages 137–143.

Zhai, J. and Barreto, A. (2006). Stress detection in computer users based on digital signal processing of noninvasive physiological variables. *Engineering in Medicine and Biology Society*, pages 1355–1358.