# BACKCHANNEL PREDICTION IN HUMAN-ROBOT INTERACTION FOR ENGAGING AGENTS

by

**Bekir Berker Türker**

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Doctor of Philosophy

in

Electrical and Electronics Engineering

KOÇ ÜNİVERSİTESİ

September 13, 2023

**BACKCHANNEL PREDICTION IN HUMAN-ROBOT INTERACTION FOR ENGAGING AGENTS**

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

**Bekir Berker Türker**

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

_____

Prof. Engin Erzin

_____

Prof. Yücel Yemez

_____

Prof. Metin Sezgin

_____

Prof. Çiğdem Eroğlu Erdem

_____

Prof. Yusuf Sahillioğlu

_____

Assist. Prof. Zafer Doğan

Date: _____

*Dedicated to my family.*

# ABSTRACT


**BACKCHANNEL PREDICTION IN HUMAN-ROBOT INTERACTION FOR ENGAGING AGENTS**

**Bekir Berker Türker**

**Doctor of Philosophy in Electrical and Electronics Engineering**

**September 13, 2023**

This thesis offers a comprehensive investigation into the realm of human-robot interaction (HRI), with a particular focus on the role of non-verbal cues such as smiles, laughs, and head nods in enhancing social engagement and naturalness. Utilizing meticulously annotated multi-modal data from naturalistic dyadic conversations, the research employs advanced machine learning techniques, including Time Delay Neural Networks (TDNNs), Support Vector Machines (SVMs), Long Short-Term Memory networks (LSTMs), and Transformer models. The work rigorously explores the utility of facial information, head movement, and audio features for the continuous detection of laughter, addressing the class imbalance problem through the effective incorporation of bagging techniques. It also delves into the impact of laughter perception and response on engagement in HRI, evaluated through objective and subjective measures in experimental setups featuring robots with laughter-responsive and non-responsive modes. Furthermore, the research presents an audio-visual prediction framework for head-nod and turn-taking events, trained and evaluated on human-human conversational datasets. A comparative approach is employed, using LSTMs as a baseline and Transformer models with cross attention mechanisms as the main proposed method, demonstrating significant improvements in the prediction performance of upcoming candidate smiles and laughs as backchannels. Collectively, these findings significantly advance our understanding of dialog management systems, offering crucial insights into the mechanics of social engagement in HRI and laying a robust foundation for future research.

# ÖZETÇE

## İNSAN-ROBOT ETKİLEŞİMİNDE İLGİ DÜZEYİNİN İYİLEŞTİRİLMESİNE YÖNELİK ARKA-KANAL SİNYAL KESTİRİMİ

**Bekir Berker Türker**
**Elektrik ve Elektronik Mühendisliği, Doktora**
**13 Eylül 2023**

Bu tez, sosyal etkileşim ve doğallığı artırmada gülümsemeler, gülmeler ve baş sallamalar gibi sözel olmayan ipuçlarının oynadığı rollere özel bir odakla, insan-robot etkileşimi (İRE) alanında kapsamlı bir araştırma sunmaktadır. Araştırma, doğal ikili konuşmalardan el ile etiketlenmiş çok kipli verileri kullanarak Zaman Gecikmeli Sinir Ağları, Destek Vektör Makineleri, Uzun Kısa Vadeli Bellek ağları ve Dönüştürücü modelleri gibi ileri makine öğrenimi tekniklerini kullanmaktadır. Çalışma, gülme tespiti için yüz ifadesi takip bilgileri, baş hareketi ve ses özelliklerinin ortaya koyduğu faydayı incelemekte ve veri sınıfı dengesizliği sorununu etkili bir şekilde çözmek için torbalama gibi teknikleri entegre etmektedir. Ayrıca, İRE'de gülme algısı ve yanıtının etkileşim üzerindeki etkisini derinlemesine incelemekte, gülme-duyarlı ve gülme-duyarsız modlara sahip robotlarla yapılan deneysel çalışmalarda objektif ve subjektif ölçümlerle etkileşimleri değerlendirmektedir. Araştırma ayrıca, insan-insan konuşma verileri üzerinde eğitilmiş ve değerlendirilmiş baş sallama ve söz sırası alma olayları için bir görsel-işitsel tahminleme çerçevesi sunmaktadır. Karşılaştırmalı bir yaklaşım için Uzun Kısa Vadeli Bellek ağları baz model olarak kullanılmış ve esas önerilen yöntem olarak çapraz dikkat mekanizmalarına sahip Dönüştürücü modeller kullanılmıştır. Bu model, etkileşim sırasında bir arka-kanal sinyali olarak oluşturulabilecek aday gülümseme veya gülme olaylarının tahmin performansında önemli iyileştirmeler göstermektedir. Topluca, bu bulgular, diyalog yönetim sistemlerini anlamamızı önemli ölçüde ilerletmekte, ERİ'de sosyal etkileşimin mekaniği ve sözel olmayan ifadeler hakkında önemli içgörüler sunmakta ve gelecekteki araştırmalar için bir temel atmaktadır.

# ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisors, Prof. Engin Erzin and Prof. Yücel Yemez, for their invaluable guidance, support, and mentorship throughout the course of this research. Their insights have been vital for shaping this work.

I am grateful to Prof. Metin Sezgin for giving me the opportunity to work and collaborate on his extraordinary projects and his guidance on almost all the necessary topics.

I am also thankful to my committee members, Prof. Çiğdem Eroğlu Erdem, Prof. Yusuf Sahillioğlu, Asst. Prof. Zafer Doğan for their thoughtful feedback and constructive critiques.

My appreciation extends to the MVGL and IUI labs for providing the resources and environment for high-quality research. Special thanks go to my colleague Zana Buçinca, whose support was very helpful and envisaged in our joint studies.

Last but not least, I wish to thank my parents, my wife, my brother and his family for their endless love, support, and encouragement throughout this challenging yet rewarding journey.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# ABBREVIATIONS

| | |
|---|---|
| AUC | Area Under the Curve |
| CNN | Convolutional Neural Networks |
| HCI | Human-Computer Interaction |
| HRI | Human-Robot Interaction |
| LSTM | Long Short-Term Memory |
| MFCC | Mel-frequency cepstrum coefficients |
| PCA | Principal Component Analysis |
| RMSE | Root Mean Squared Error |
| RNN | Recurrent Neural Network |
| ROC | Reciever Operating Characteristic |
| SVM | Support Vector Machines |
| TDNN | Time-Delay Neural Network |

# Chapter 1

# INTRODUCTION

## *1.1 Motivation*

Human-robot interaction (HRI) has witnessed significant attention from both academic and industrial sectors, creating an imperative need to investigate the subtleties of non-verbal cues like smiles and laughter. Seminal works have previously underscored the importance of non-verbal communication in human-human interactions. For example, [Mehrabian and Epstein, 1972] suggested that as much as 93% of emotional meaning in conversations could be conveyed through non-verbal cues. This foundational understanding raises an immediate question—how can robots be designed to capture this nuance and engage meaningfully with humans?

In this thesis, chapter 2 draws inspiration from studies like [Krikke and Truong, 2013, Kennedy and Ellis, 2004, Knox and Mirghafori, 2007, Truong and Leeuwen, 2005], highlighting the challenges and importance of laughter detection in social interactions. These studies often employ machine learning techniques, but the challenge of continuous laughter detection in naturalistic settings has often been overlooked. Our work fills this research gap by introducing a meticulously annotated audio-facial database. The chapter goes beyond conventional feature selection techniques, leveraging mutual information-based criteria to identify discriminative features. Furthermore, the use of bagging to address class imbalance echoes work like [Chawla et al., 2002], providing a balanced approach to the problem.

In Chapter 3, we build upon existing literature investigating the role of responsive behaviors in robots. Previous studies like [Sidner et al., 2005a, Peters et al., 2009, Glas and Pelachaud, 2015] have delved into the importance of engagement in HRI,

but few have directly evaluated how laughter could act as a significant backchannel [Pecune et al., 2015, Niewiadomski et al., 2013b, El Haddad et al., 2016]. Our research adds an empirical dimension to these theoretical constructs, evaluating human engagement with a laughter-responsive robot compared to a non-responsive one.

Chapter 4 leverages the foundational research on turn-taking and head nods in dyadic interactions, often highlighted in works like [Skantze et al., 2014, Skantze, 2021, Jokinen et al., 2013, Calisgan et al., 2012, Chang et al., 2022]. Using the IEMOCAP dataset, we apply machine learning algorithms like SVM and LSTM to predict these cues, thereby contributing to the body of research that seeks to make HRI more dynamic and natural.

The fifth chapter incorporates recent advances in machine learning and multimodality to tackle the prediction of smiles/laughs as backchannels. Studies such as [Vaswani et al., 2017] and [Hochreiter and Schmidhuber, 1997a] have revolutionized the field of sequence modeling by introducing Transformer models and LSTMs, respectively. Studies like [Xie et al., 2019, Ma et al., 2019, Zou et al., 2022, Truong et al., 2021, Zhong et al., 2019, Li et al., 2020] have been using the advancements successfully. Our work builds upon these contributions by employing them in the unique context of HRI, enabling more engaging and natural social robots.

## *1.2 Contributions and Organization*

This thesis introduces a framework for predicting backchannels in HRI, specifically focusing on smiles and laughter. The overarching aim is to augment robots' engagement and naturalness during interactions. To achieve this, the research journey encompasses multiple facets: from the detection of human laughter to the study of engagement levels with a laughter-responsive robot and from the prediction of head-nods and turn-taking events in human-human interactions for potential application in HRI to the employment of state-of-the-art deep learning techniques for the prediction of smiles and laughter. Subsequent sections describe the contributions and organization of the thesis.

Chapter 2 addresses the challenge of continuous laughter detection in naturalistic conversations, utilizing a rigorously annotated audio-facial database. Machine learning classifiers, particularly Time Delay Neural Networks (TDNNs) and Support Vector Machines (SVMs), are employed to dissect the utility of distinct facial features, head movements, and audio cues. The chapter also tackles the class imbalance problem through bagging, a resampling technique that improves classifier performance.

Chapter 3 builds on these foundational elements, investigating how a robot's responsiveness to laughter impacts human engagement. In an experimental setup featuring a back-projected robot head, the robot alternates between two operational modes, laughter-responsive and laughter-non-responsive, to gauge the consequent variations in user engagement. This approach allows us to empirically validate the critical role of laughter as a backchannel in HRI, offering both objective and subjective metrics to assess interaction quality.

In Chapter 4, we further enrich the landscape of non-verbal cues by introducing head nods and turn-taking as significant elements in dyadic conversations. Utilizing the IEMOCAP dataset, Support Vector Machines (SVMs) and Long Short-Term Memory (LSTM) are trained for predictive modeling. The chapter provides a comprehensive analysis of unimodal and multi-modal classification performance, illuminating the importance of these cues in conversation dynamics.

Chapter 5, which forms the crux of this thesis, employs advanced deep-learning techniques to predict smiles as backchannels in dyadic interactions. Long Short-Term Memory (LSTM) networks serve as the baseline model, and Transformer models equipped with cross-attention mechanisms are introduced as an advanced alternative. Our results unequivocally indicate that Transformer models, when configured with optimized acoustic and temporal windows, significantly outperform traditional LSTM networks.

### *1.3 List of Publications*

- Turker, B. B., Erzin, E., Yemez, Y., & Sezgin, T. M. (2018). Audio-Visual Prediction of Head-Nod and Turn-Taking Events in Dyadic Interactions. Interspeech, 1741–1745.

- Devillers, L., Rosset, S., Duplessis, G. D., Bechade, L., Yemez, Y., Turker, B. B., ... Others. (2018). Multifaceted engagement in social interaction with a machine: The joker project. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), 697–701. IEEE.

- Turker, B. B., Yemez, Y., Sezgin, T. M., & Erzin, E. (2017). Audio-facial laughter detection in naturalistic dyadic conversations. IEEE Transactions on Affective Computing, 8(4), 534–545.

- Turker, B. B., Buçinca, Z., Erzin, E., Yemez, Y., & Sezgin, T. M. (2017). Analysis of Engagement and User Experience with a Laughter Responsive Social Robot. Interspeech, 844–848.

- Devillers, L., Rosset, S., Duplessis, G. D., Sehili, M. A., Béchade, L., Delaborde, A., ... Others. (2015). Multimodal data collection of human-robot humorous interactions in the joker project. 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), 348–354. IEEE.

# Chapter 2

# MULTIMODAL LAUGHTER DETECTION

## *2.1 Introduction*

Laughter serves as an expressive social signal in human communication and conveys distinctive information on the affective state of conversational partners. As affective computing is becoming an integral aspect of human-computer interaction (HCI) systems, automatic laughter detection is one of the key tasks towards designing more natural and human-centered interfaces with better user engagement [Devillers et al., 2015].

Laughter is primarily a nonverbal vocalization accompanied by body and facial movements [Ruch and Ekman, 2001]. Most of the existing automatic laughter detection methods in the literature have focused on audio-only information [Petridis and Pantic, 2011, Cosentino et al., 2016]. This is mainly because audio is relatively easier to capture and analyze than other modalities of laughter and often alone sufficient for humans to identify laughter. Yet visual cues due to accompanying body and facial motion also help humans to detect laughter, especially in the presence of cross-talk, environmental noise, and multiple speakers. While there is some recent trend in the community for automatic laughter detection from full body movements [Niewiadomski et al., 2016, Griffin et al., 2015], there are so far few works that exploit facial motion [Petridis and Pantic, 2011]. The main bottleneck here, especially for incorporating facial data, is the lack of multimodal databases from which facial laughter motion can reliably be extracted. The existing works that incorporate facial motion mostly make use of audiovisual recordings. Hence, they rely on facial feature points extracted automatically from video data, which are, in fact, difficult to reliably track in the case of sudden and abrupt facial movements such as in laugh-

ter. As a result, the common practice is to use the resulting displacements of facial feature points (in the form of FAPS parameters, for instance) as they are without further analysis. In this respect, the primary goal of this work is to investigate facial and head movements for their use in laughter detection over a multimodal database comprising facial 3D motion capture data and audio.

We address several challenges involved in the automatic detection of laughter. First, we present detailed annotation of laughter segments over an audio-facial database comprising explicit 3D facial mocap data. Our annotation includes cross-talks and environmental noise as well. Second, using this annotated database, we investigate different ways of incorporating facial information and head movement to boost laughter detection performance. In particular, we focus on discriminative analysis of facial features contributing to laughter and perform feature selection based on mutual information. Another issue that we consider is the relative scarcity of laughter instances in real-world conversations, which hinders the machine-learning task due to highly imbalanced training data. To address this problem, we incorporate bagging into our classification pipeline to better model non-laughter audio and motion. Finally, we analyze the performance of the proposed multimodal fusion setup that uses selected combinations of audio-facial features and classifiers for continuous detection of laughter in the presence of cross-talks and environmental noise.

The work presented in this chapter falls under the general field of affective computing. However, laughter detection is very different from affect recognition [Zeng et al., 2009] that has been receiving the main thrust in the community. As indicated by various active strands of work, laughter detection is an entirely separate field of interest driven by several research groups [Truong and Leeuwen, 2007, Laskowski and Schultz, 2008, Petridis and Pantic, 2011, Reuderink et al., 2008, Petridis et al., 2013a, Petridis et al., 2011]. Laughter falls under the category of 'affect bursts' which denote *specific identifiable events* [Scherer, 1994]. This is unlike the general work in affect recognition, which attempts to *continuously assess the emotional state* of the person of interest.

The work on laughter detection can be categorized into two major lines: unimodal and multimodal approaches. Unimodal approaches have mainly explored the audio modality. For example, Truong et al. have focused on laughter detection in speech [Truong and Leeuwen, 2007], and Laskowski et al. explored laughter detection in the context of meetings [Laskowski and Schultz, 2008]. Other unimodal work focused on body movements [Griffin et al., 2015, Niewiadomski et al., 2016]. Griffin et-al. [Griffin et al., 2015] studied how humans perceive laughter through avatar-based perceptual studies and also explored automatic recognition of laughter from body movements. In another work, Griffin et al. [Griffin et al., 2013] presented recognition (not detection) results on pre-segmented laughter instances falling into five different categories. In contrast, here, we take a multimodal approach and perform detection. We use the term "detection" to refer to the task of segmentation and classification over a continuous data stream and the term "recognition" for the task of classification on pre-segmented samples.

In another work, Niewiadomski et-al. [Niewiadomski et al., 2016] identified sets of useful features for discriminating laughter and non-laughter segments. They used discriminative and generative models for recognition and showed that automatic recognition through body movement information compares favorably to human performance.

To distinguish pre-segmented non-verbal vocalizations using audio-only features, solutions employing different learning methods have been proposed. Schuller et al. [Schuller et al., 2008] used a variety of classifiers on dynamic and static representations to differentiate non-verbal vocalizations (laughter, breathing, hesitation, and consent). Among the various classifiers they used, hidden Markov models (HMMs) outperformed other classifiers, such as hidden conditional random fields (HCRFs) and support vector machines (SVM). Unlike what we present, this work is unimodal in nature. Our work complements these lines of work by shedding more light on how audio and motion information contribute to laughter recognition as individual modalities.

Other lines of work have indirectly studied laughter detection while addressing

other problems. For example, since laughter is treated as noise in speech recognition, its detection, segmentation, and suppression have received attention [Prylipko et al., 2012]. Here, laughter detection is our primary concern, and we treat it rigorously.

Our work is more closely related to the multimodal laughter detection and recognition systems [Escalera et al., 2009, Petridis and Pantic, 2011, Reuderink et al., 2008, Petridis et al., 2013a, Petridis and Pantic, 2008, Ito et al., 2005]. Escalera et-al. combined audio information with smile-laughter events detected at the frame level and identified regions of laughter [Escalera et al., 2009]. Petridis et-al. proposed methods for discrimination between pre-segmented laughter and speech using both audio and video streams [Petridis and Pantic, 2011, Petridis and Pantic, 2008]. The features extracted are facial expressions and head pose from video, and cepstral and prosodic features from audio. They have also shown that the decision-level fusion of the modalities outperformed audio-only and video-only classifications using decision rules as simple as the SUM rule. Recently Petridis et al. [Petridis et al., 2013a] have proposed a method for laughter detection using time delay neural networks (TDNN). They have explained their relatively lower values for precision, recall, and $F_1$ score by the presence of imbalanced data, where a typical stream would have way more non-laughter frames than laughter ones. Our work further supports the findings of these studies and also gives a clear advantage through the use of bagging to deal with imbalanced databases.

In other multimodal work, Cosker and Edge [Cosker and Edge, 2009] presents an analysis of the correlation between voice and facial marker points using HMMs in four non-speech articulations: laughing, crying, sneezing, and yawning. Although their work is geared towards synthesizing facial movements using sound, it also provides useful insights into laughter recognition. A recent study by Krumhuber and Scherer [Krumhuber and Scherer, 2011] shows that the facial action units, coded using the Facial Action Coding System (FACS), exhibit significant variations for different affect bursts and hence can serve as cues in detecting and recognizing laughter.

Scherer et-al. [Scherer et al., 2012] proposed a multimodal laughter detection sys-

tem based on Support Vector Machines, Echo State Networks, and Hidden Markov Models. Although they use body and head movement information, it is extracted from video. Similarly, Reuderink et al. [Reuderink et al., 2008] perform audiovisual laughter recognition on a modified and re-annotated version of the AMI Meeting Corpus [Carletta et al., 2006]. They report performance measures over data containing 60 randomly selected laughter and 120 non-laughter segments. They use 20 points tracked on the face to capture the movements in the video. Both of these approaches use video sequences for motion and facial feature point extraction in contrast with the use of mocap data in our work.

In another recent work, Turker et al. proposed a method for recognition between pre-segmented types of affect bursts, namely, laughter and breathing, using HMMs [Türker et al., 2014]. Although this method is promising, it could not be used directly for the continuous detection of laughter over input streams.

The data we use in our evaluation is a broadened version of the IEMOCAP database extended through a painstaking annotation effort and serves as one of our main contributions. Although there are databases of unimodal and multimodal audiovisual laughter data [Suarez et al., 2012, Petridis et al., 2013b, McKeown et al., 2012, Niewiadomski et al., 2013c], our database stands out by the fact that it comprises explicit facial mocap data and has been annotated for cross-talk and environmental noise. Hence, we could train models with training data selected based on their clean, noisy, and/or cross-talk labels.

### *2.1.1 Contributions*

In view of the related work discussed previously, the contributions of this work can be summarized as follows:

- We provide detailed annotation of laughter segments over an existing audio-facial database that comprises 3D facial mocap data and audio, considering cross-talks and environmental noise.

- We perform a discriminative analysis on facial features via feature selection

based on mutual information to determine facial movements that are most relevant to laughter over the annotated database.

- We construct a multimodal laughter detection system that compares favorably to state-of-the-art, especially the facial laughter detection performs outstanding among the best-performing methods in the literature.

- We analyze the performance for continuous detection of laughter and demonstrate the advantage of incorporating facial and head features, especially to handle cross-talk and environmental noise.

## *2.2 Database*

Our analysis and experimental evaluations used the interactive emotional dyadic motion capture database (IEMOCAP) to study expressive human interactions [Busso et al., 2008]. The IEMOCAP is an audio-facial database that provides motion capture data of face, head, and partially hands and speech. The corpus has five sessions with ten professional actors participating in dyadic interactions. Each session comprises spontaneous conversations under eight hypothetical scenarios and three scripted plays to elicit rich emotional reactions. The recordings of each session are split into clips, where the total number of clips is 150 with a total duration of approximately 8 hours.

The focus of the IEMOCAP database is to capture emotionally expressive conversations. The database is not specifically intended to collect laughter; hence, the laughter occurrences in the database are not pre-planned in any recorded interactions, whether spontaneous or scripted; they are generated by the actors based on the emotional content and flow of the conversation. Instead of reading directly from a text, the actors rehearse their roles in advance (in the case of scripted plays) or improvise emotional conversations (in the case of spontaneous interactions). The database, in this sense, falls under the category of "semi-natural" according to the taxonomy given in [Douglas-Cowie. et al., 2003]. The genuineness of acted emotions is an open issue in most of the existing "semi-natural" databases. Yet several works

in the literature address this issue and suggest possible strategies to increase the genuineness of acted emotions [Douglas-Cowie. et al., 2003, Busso and Narayanan, 2008, Bryant and Aktipis, 2014]. In this sense, IEMOCAP can be viewed as an effort to capture naturalistic multimodal emotional data through dyadic conversations performed by professional actors under carefully designed scenarios. We note that there exists yet no emotional database that includes accurate facial measurements in the case of fully natural laughter.

In the IEMOCAP database, a VICON motion capture system tracks 53 face markers, 2 head markers, and 6 hand markers of the actors at 120 frames per second. The placement of the facial markers is consistent with the feature points defined in the MPEG-4 standard. In each session, only one actor has markers. We call the actor with markers as the speaker and the other actor as the listener. Speakers' data will be the main focus of this study, as they include audio and facial marker information. However, listeners also impact the audio channel by creating a cross-talk effect on speakers' laughter. The authors of [Busso et al., 2008] note that the markers attached to speakers during motion data acquisition are very small so as not to interfere with natural speech. According to [Busso et al., 2008], the subjects also confirmed that they were comfortable with the markers, which did not prevent them from speaking naturally.

Throughout this study, the motion capture data of the facial and head markers will be referred to as motion features. The proposed laughter detection will use audio and motion features in a multimodal framework.

### *2.2.1 Laughter Annotation*

Although the text transcriptions were available with the IEMOCAP, the laughter annotations were missing. We have performed a careful and detailed laughter annotation task over the full extent of the database. The annotation effort has been carried out with one annotator. Laughter segments are identified with a clear presence of audiovisual laughter events. The laughter annotations have been performed only for the speaker over each clip in the database. In addition, the speech activity

of the listener has also been annotated around the laughter segments of the speaker, which can be defined as cross-talk for the laughter event. Similarly, the acoustic noise appears as a disturbance to laughter events. We define noise as any distinguishable environmental noise in audio caused by some external factors, such as the footsteps of the recording crew and the creaking chairs of participants (that especially happens when laughing due to abrupt body movements). We have annotated the presence of environmental noise in speakers' and listeners' recordings around the laughter segments of speakers. Note that our annotation does not include the noise due to data acquisition. The speech activity and noise presence annotations allow us to label the laughter conditions of a speaker as clean, noisy, and/or cross-talk. A sample annotation stream is shown in Figure 2.1.

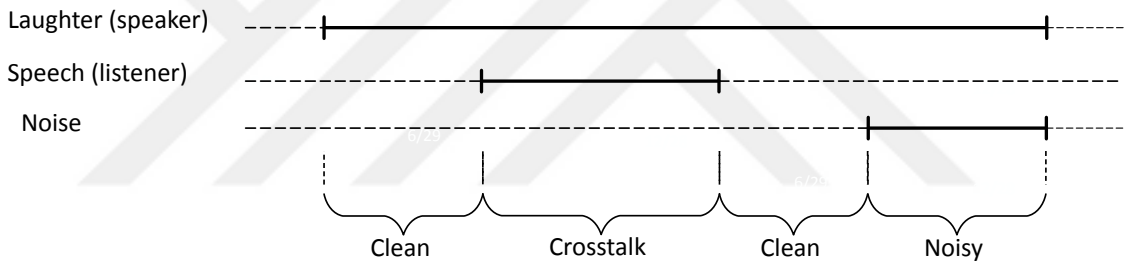



Figure 2.1: Sample annotation fragment with speaker laughter, listener speech, and noise.

Table 2.1: Laughter annotation statistics

| | **Clean** | **Noisy** | **Cross-talk** | **All** |
|---|---|---|---|---|
| **Number of occurrences** | 231 | 44 | 194 | 248 |
| **Total duration (sec)** | 213.92 | 17.49 | 165.25 | 382.00 |
| **Average duration (sec)** | 0.93 | 0.40 | 0.85 | 1.54 |

All five sessions of the IEMOCAP database have been annotated as described above. Table 2.1 shows a summary of the laughter annotations. The first three columns present the number of occurrences and the duration information for clean,

noisy, and cross-talk conditions. The last column of the table is a summary of the laughter annotation of speakers before pruning the noisy and cross-talk conditions. After pruning the noisy and cross-talk segments, the total duration of laughter segments decreases from 382.00 sec to 213.92 sec. The average duration of the laughter segments also decreases from 1.54 sec to 0.93 sec. This is due to the pruning process, which splits some long laughter segments into shorter ones. Note that we try to keep as much clean laughter data as possible to use them in model training.

As given in Table 2.1, the total duration of clean laughter sequences is 213.92 sec, while the annotated IEMOCAP database is 8 hours. This causes a data imbalance between the two event classes of interest, laughter and non-laughter. We construct a balanced database (BLD) to train our classifiers, which includes all clean laughter segments and a subset of non-laughter audio segments from IEMOCAP, excluding all noisy and cross-talk laughter segments. The non-laughter samples, which define a reject class for the laughter detection problem, are picked randomly to match the total duration of laughter.

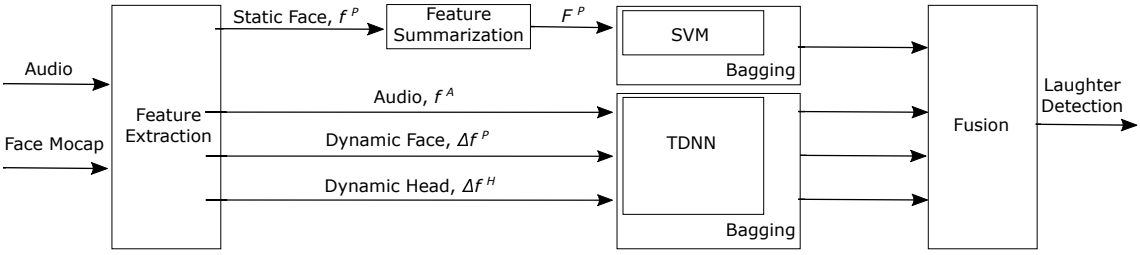



Figure 2.2: Block diagram of the proposed laughter detection system. The diagram shows the best-performing setup for classifier-feature combinations.

## *2.3 Methodology*

The main objective of this work is to detect laughter segments in naturalistic dyadic interactions using audio, head, and facial motion. The block diagram of the proposed system is illustrated in Figure 2.2. Audio and face motion capture data are inputs

to the system, from which short-term audio and motion features are extracted. Audio is represented with mel-frequency cepstral coefficients (MFCCs) and prosody features, whereas motion features are extracted from 3D facial points and head tracking data in the form of positions, displacements, and angles. Two different types of classifiers, i.e., support vector machines (SVM) and time-delay neural networks (TDNN), receive a temporal window of short-term features or their summarizations to perform laughter vs. non-laughter binary classification. The classification task is repeated every 250 msec over overlapping temporal windows of length 750 msec. While the TDNN classifier works on short-term features, the SVM classifier runs on statistical summarization of them. Both types of classifiers make use of bagging to better handle data imbalance between laughter and non-laughter classes. Finally, different classifier-representation combinations are integrated via decision fusion to perform laughter detection on each temporal window.

### *2.3.1 Audio Features*

Spectral properties and prosodic structures can characterize acoustic laughter signals. The mel-frequency cepstral coefficient (MFCC) representation is the most widely used spectral feature in speech and audio processing, and it was successfully used before in characterizing laughter [Petridis and Pantic, 2011]. We compute 12-dimensional MFCC features using a 25 msec sliding Hamming window at intervals of 10 msec. We also include the log energy and the first-order time derivatives in the feature vector. The resulting 26-dimensional spectral feature vector is represented with $f^M$.

Prosody characteristics at the acoustic level, including intonation, rhythm, and intensity patterns, carry important temporal and structural clues for laughter. We choose to include speech intensity, pitch, and confidence-to-pitch into the prosody feature vector as in [Bozkurt et al., 2015, Bozkurt et al., 2016a]. Speech intensity is extracted as the logarithm of the average signal energy over the analysis window. Pitch is extracted using the YIN fundamental frequency estimator, which is a well-known auto-correlation-based method [de Cheveigne¥ and Kawahara, 2002].

Confidence-to-pitch delivers an auto-correlation score for the fundamental frequency [Bozkurt et al., 2016a].

Since prosody is speaker and utterance-dependent, we apply mean and variance normalization to prosody features. The mean and variance normalization of prosody features is performed over small connected windows of voiced articulation, which exhibits pitch periodicity. Then, the normalized intensity, pitch, confidence to pitch features, and the first temporal derivative of these three parameters are used to define the 6-dimensional prosody feature vector denoted by $f^S$. The extended 32-dimensional audio feature is then obtained by concatenating spectral and prosody feature vectors: $f^A = [f^M f^S]$.

### *2.3.2 Head and Facial Motion Features*

We represent the head pose using a 6-dimensional feature vector $f^H$ that includes $x$, $y$, $z$ coordinates of the head position, and the Euler angles representing head orientation with respect to three coordinate axes. The reference point for the head position and the three coordinate axes are common to all speakers and computed from face instances with neutral pose [Busso et al., 2008]. We will refer to $f^H$ as static head features, whereas dynamic head features will be represented by $\Delta f^H$, which are simply the first derivatives of static features. We note that dynamic head features are less dependent on global head pose and carry more explicit information about head movements that are discriminative for laughter detection, such as nodding up and down. Note also that the few methods existing in the literature that incorporate explicit 3D head motion for laughter detection [Niewiadomski et al., 2016, Griffin et al., 2015] calculate head-related features based on the positioning of the head with respect to the full body such as the distance between shoulders and head as in [Niewiadomski et al., 2016]. However, these features are not applicable when dealing with audio-facial data, which does not include full body measurements.

Likewise, we define two sets of facial features: static and dynamic. The static facial feature vector is obtained by concatenating the 3D coordinates of the tracked facial points. Hence, the dimension of this vector is $3 \times M$, where $M$ is the number

of facial points, which is at most 53 in our case. We assume that scale is normalized across all speakers and the 3D coordinates are given with respect to a common origin, which is the tip of the nose in the IEMOCAP database [Busso et al., 2008]. We denote the static facial feature vector by $f^P$ and the dynamic facial feature vector by $\Delta f^P$, which is the first derivative of the static version. We note that pose invariance is less problematic in this case than head motion since facial points are compensated for rigid motion.

### *2.3.3 Feature Summarization*

We use feature summarization for the temporal characterization of laughter. We compute a set of statistical quantities describing each frame-level feature's short-term distribution over a given window for feature summarization. These quantities comprise 11 functionals, more specifically mean, standard deviation, skewness, kurtosis, range, minimum, maximum, first quantile, third quantile, median quantile, and inter-quantile range, which were successfully used before by Metallinou et al. [Metallinou et al., 2013] for continuous tracking of emotional states from speech and body motion. We denote the window-level statistical features resulting from the summarization of frame-level features $f$ by $F$. Hence, the statistical features computed on audio, static and dynamic head and facial features are represented by $F^A$, $F^H$, $\Delta F^H$, $F^P$, and $\Delta F^P$, respectively. The dimension of each of these statistical feature vectors can be calculated as 11 times the dimension of the corresponding frame-level feature vector. We will later use these window-level statistical features to feed SVM classifiers for laughter detection.

### *2.3.4 Discriminative Analysis of Facial Laughter*

In this subsection, we perform discriminative analysis on facial points to determine the relevance of each point in the formation of laughter expressions over the given database. We also explore possible correlations between facial points to eliminate redundancies in distinguishing laughter. Such correlations exist especially between symmetrical facial points (e.g., right cheek vs. left cheek) and between points be-

longing to a muscle group. We will later use the results of this analysis to define optimal sets of facial features to be fed into our classification pipeline for laughter detection.

We employ the feature selection method mRMR (minimum Redundancy Maximum Relevance)[Peng et al., 2005]. This method assigns an importance score to each feature, which measures its relevance to target classes under minimum redundancy conditions. Relevance is defined based on mutual information between features and the corresponding class labels (laughter vs. non-laughter in our case), whereas redundancy takes into account the mutual information between features. Hence, when features are sorted in a list with respect to importance in descending order, the first $m$ features from this list form the optimal $m$-dimensional feature vector that carries the most discriminative information for classification. Another useful feature selection method that we utilize is called maxRel (Maximum Relevance)[Peng et al., 2005], which ranks features without imposing any redundancy condition and takes only relevance into account when assigning importance to features. We report the results using both methods, maxRel to observe the importance of individual points for laughter, and mRMR to find optimal subsets of features to be fed into our classification engine.

We extract window-level statistical features from a realization of the BLD with their class labels and apply mRMR and maxRel methods. Both methods result in a feature ranking list. Since each facial marker point on the face has $3 \times 11$ statistical features in our case, the complete list has $53 \times 3 \times 11$ features, where 53 is the number of facial markers. We employ a voting technique to quantify the importance of each facial point based on this ordered list of features with individual scores. Each point collects votes from 33 contributor features, where each contributor votes in proportion to its importance score resulting from maxRel or mRMR. The accumulated votes finally sum up to an overall importance score for each facial point.

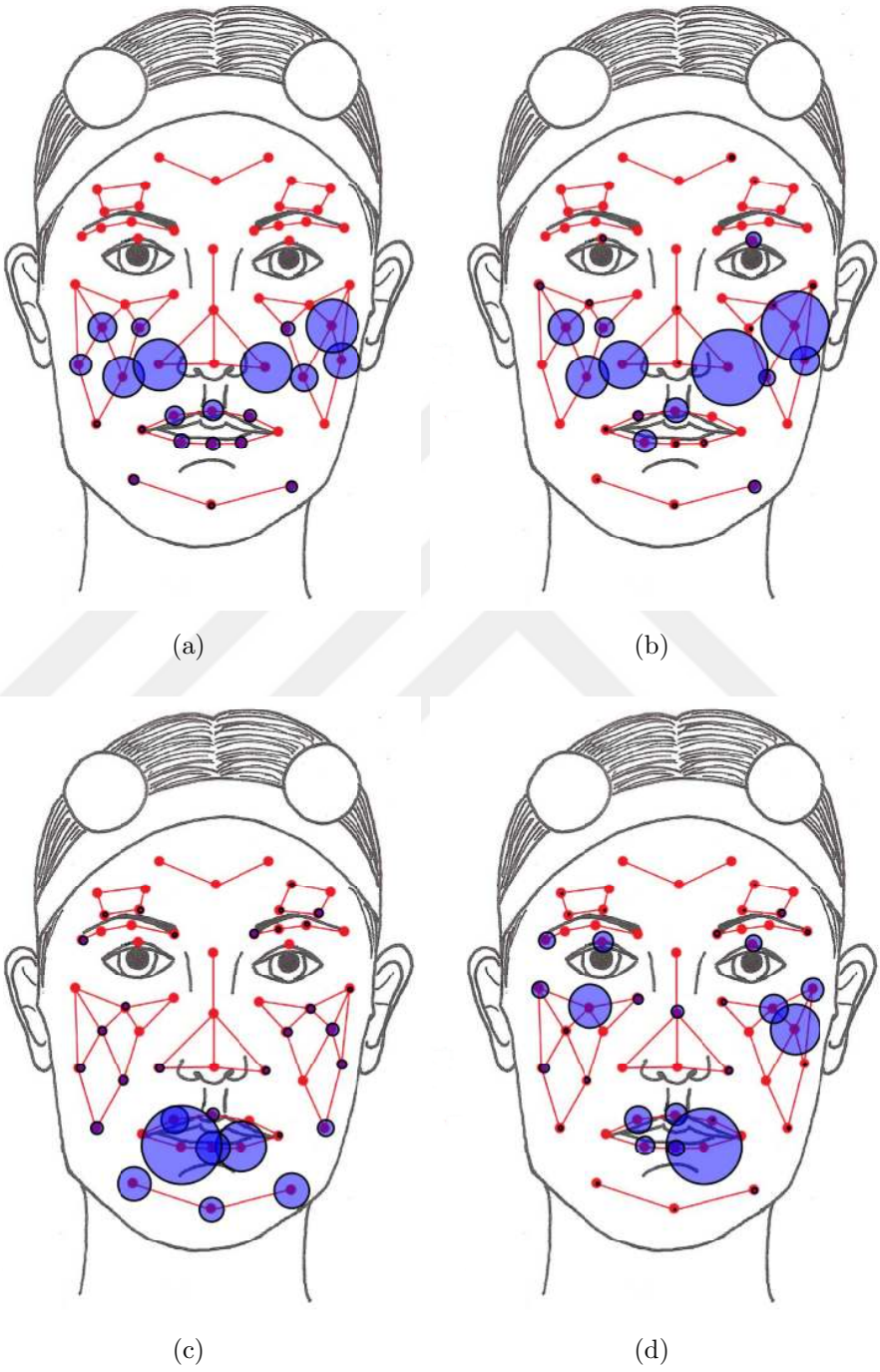
For visualization, the importance scores resulting from the underlying selection process are used to modulate the radius of a disc around each marker point. Figures 2.3a and 2.3b display the results of maxRel and mRMR for static facial features

$F^P$, respectively. In these figures, the size of a disc is proportional to the importance of its center point. In the case of maxRel, as expected, we observe a more symmetric and balanced distribution of importance focused on certain facial regions, especially the mouth and cheek regions. In the case of mRMR, however, we see that the distribution is not that symmetric, and the distribution of importance tends to concentrate on fewer points.

Figures 2.3c and 2.3d display the results of maxRel and mRMR for dynamic facial features $\Delta F^P$, respectively. For dynamic features, it is evident that most of the relevance is concentrated, with a symmetric and balanced distribution, on mouth and chin regions, which have a relatively larger amount of movement than any other regions of the face. In mRMR results, however, we see that the points over the chin region no longer appear to be important. This is probably because the motion of the points on the mouth carries very similar information since lower lip points can rarely move independently from the chin. In Figure 2.3d, we also observe that points around the eyes start to get more important, indicating the genuineness of the laughter in the database [Ruch and Ekman, 2001].

### *2.3.5 Classification*

As discussed briefly at the beginning of this section, two statistical classifiers, SVM and TDNN, are employed for the laughter detection system. SVM is a binary classifier based on statistical learning theory, which maximizes the margin that separates samples from two classes [Cortes and Vapnik, 1995]. SVM projects data samples to different spaces through kernels that range from simple linear to radial basis function (RBF) [Chang and Lin, 2011]. We consider the summarized statistical features as inputs of the SVM classifier to discriminate laughter from non-laughter. On the other hand, TDNN is an artificial neural network model in which all the nodes are fully connected by directed connections [Waibel et al., 1989]. Inputs and their delayed copies construct the nodes of the TDNN, where the neural network becomes time-shift invariant and models the temporal patterns in the input sequence. We use the TDNN classifier to model the temporal structures of laughter events, as it



(a) (b)

(c) (d)

Figure 2.3: Visualization of the importance of facial points for laughter using (a) maxRel - static features, (b) mRMR - static features, (c) maxRel - dynamic features, and (d) mRMR - dynamic features. The size of a disc is proportional to the importance of its center point.

has been successfully used by Petridis et al. [Petridis et al., 2013a]. In this work, we adopt their basic TDNN structure, which has only one hidden layer. The further details of the classifier structure, its parameters, and the optimization process are explained in Section 2.4.

*Bagging*

In the nature of daily conversations, laughter occurrences, and their durations are sparse within non-laughter utterances. This data imbalance problem has been pointed out in Section 2.2.1. This problem has been addressed, and several solutions have been suggested in the literature [Galar et al., 2012]. For instance, the SVM classifier performs better when class samples are balanced in the training [Akbani et al., 2004]. Otherwise, it tends to favor the majority class. Several methods have been proposed for this problem, such as down-sampling the majority class or up-sampling the minority class by populating with noisy samples. We choose to down-sample the majority class and use the BLD database for model training as defined in Section 2.2.1. Since the BLD database includes a reduced set of randomly selected non-laughter segments, it may not represent the non-laughter class fairly well. We use bagging to compensate for this effect, where a bag of classifiers is trained on different realizations of the BLD database and combined using the product rule for the final decision [Kittler et al., 1998, Erzin et al., 2005]. Hence, in bagging, each classifier has a balanced training set, modeling the non-laughter class over different realizations. The bagging approach is expected to bring in modeling and performance improvements. We investigate the benefit of bagging in laughter detection and report performance results in Section 2.4.

*Fusion*

The last block of Figure 2.2 is multimodal fusion, where we perform decision fusion of classifiers with different feature sets. Decision fusion of multimodal classifiers is expected to reduce the overall uncertainty and increase the robustness of the laughter detection system. Suppose that $N$ different classifiers, one for each of the

$N$ feature representations $f_1, f_2, ..., f_N$, are available, and for the $n$-th classifier a 2-class log-likelihood function is defined as $\rho_n(\lambda_k)$, $k = 1, 2$, respectively for the laughter and non-laughter classes. The fusion problem is then to compute a single set of joint log-likelihood functions $\rho(\lambda_1)$ and $\rho(\lambda_2)$ over these $N$ different classifiers. The most generic way of computing joint log-likelihood functions can be expressed as a weighted summation:

$$\rho(\lambda_k) = \sum_{n=1}^{N} \omega_n \rho_n(\lambda_k) \quad \text{for } k = 1, 2, \tag{2.1}$$

where $\omega_n$ denotes the weighting coefficient for classifier $n$, such that $\sum_n \omega_n = 1$. Note that when $\omega_n = \frac{1}{N}$ $\forall n$, (2.1) is equivalent to the product rule [Kittler et al., 1998, Erzin et al., 2005]. In this study, we employ the product rule and set all weights equal for decision fusion of the multimodal classifiers.

## *2.4 Experimental Results*

Experimental evaluations are performed across all modalities, including audio features and static/dynamic motion features, using SVM and TDNN classifiers. All laughter detection experiments are conducted in a leave-one-session-out fashion, which results in a 5-fold train/test performance analysis. Since subjects differ across sessions, the reported results are also speaker-independent. In each fold, training is carried out over a realization of the BLD extracted from the four training sessions. Hence, the training set is balanced and does not include noisy and cross-talk samples. In the testing phase, laughter detection is carried out over the whole test session data, including all non-laughter segments and all laughter conditions. As discussed in Section 2.3, laughter detection is performed as a classification task for every 250 msec over temporal windows of length 750 msec. Any temporal window that contains a laughter segment longer than 375 msec, is taken as a laughter event.

The RBF kernel is used for all modalities with the hyper-parameters $c$ and $\gamma$ for the SVM classifiers. In order to set the hyper-parameters, we execute independent 4-fold leave-one-session-out validation experiments for each fold of the 5-fold

train/test. In these validation experiments, the classification performance is evaluated on a grid of hyper-parameter values. Finally, we set the $c$ and $\gamma$ parameters to maximize the classification performance. Note that this procedure yields different parameter settings for each fold of the 5-fold train/test evaluation, but on the other hand, it ensures independence of the parameter setting procedure from the test data.

The TDNN classifiers are defined by two parameters: the number of hidden nodes and the time delay. Since TDNN classifiers exhibit increasing performance as the number of hidden nodes and time delay increase, we tend to set these parameters as low as possible to minimize the computational complexity (especially the training phase) while maintaining a high classification performance.

Finally, the resulting laughter detection performances are reported in terms of the area under the curve (AUC) percentage of the receiver operating characteristic (ROC) curves, which represents the overall performance over all operating points of the classifier [Fawcett, 2006]. Note that AUC is 100% for the ideal classifier and 50% for a random classifier.

### *2.4.1 Results on Audio*

The SVM and TDNN classifiers are considered for audio laughter detection experiments. In the case of SVM, the summarized audio features, $F^M$ and $F^A$, are used. Recall that $F^M$ includes spectral features, whereas $F^A$ includes both prosodic and spectral features. Probabilistic outputs of the SVM classifier are obtained, and then the ROC curves are calculated. In TDNN, a single hidden layer with 30 nodes using frame-level features, $f^M$ and $f^A$, is employed, and likelihood scores of the laughter and non-laughter classes are produced by the output layer, which has two nodes. The time-delay parameter is used as four as in [Petridis et al., 2013a]. We also apply 10-fold bagging to the best-performing classifiers.

Figure 2.4 plots all ROC curves and reports AUC performances for six classifier-feature combinations. We observe that the ROC curves are clustered into two groups, where the SVM classifiers without bagging constitute the first group with lower performances. Note that, bagging helps more to the SVM classifier with 3% AUC

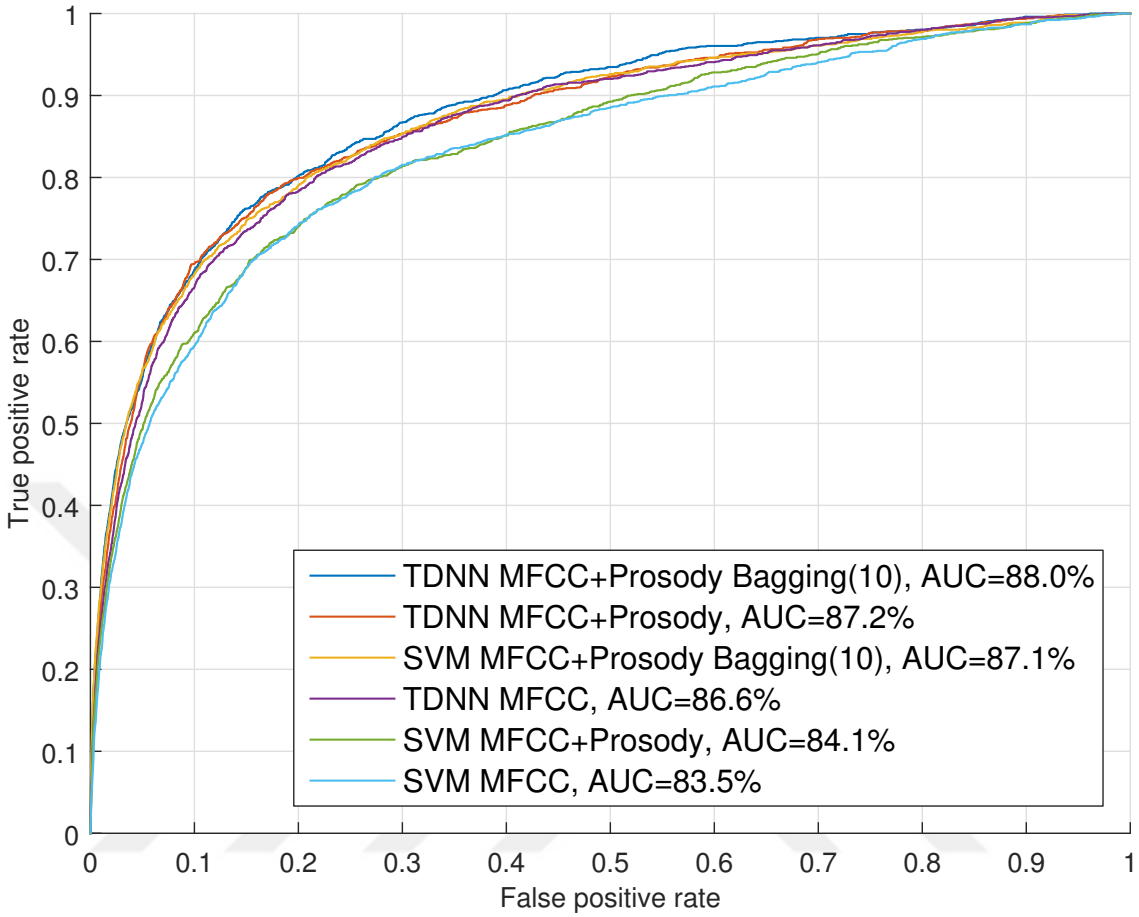



Figure 2.4: Audio laughter detection ROC curves and AUC performances of the SVM and TDNN classifiers with different feature sets and with/without bagging.

performance improvement, whereas the TDNN classifier improves 0.8% with bagging. The best performing classifier-feature combination is observed as the TDNN classifier with bagging using the audio feature $f^A$, which achieves 88.0% AUC performance.

### *2.4.2 Results on Head and Facial Motion*

Laughter detection from head and facial motion features is performed by using SVM and TDNN classifiers. The parameters of the TDNN classifier, number of hidden nodes ($n_h$) and time-delay ($\tau_d$), are set for the motion features by testing a fixed number of configurations with $\tau_d = 4, 8, 16$ and $n_h = 5, 10, 20$.

In our preliminary studies, the static head features, $f^H$ and $F^H$, have performed poorly for laughter detection, possibly because these features have severe pose invariance problems. Hence, in this work, we consider only the dynamic features $\Delta f^H$ and $\Delta F^H$ for head motion representation. The laughter detection performances of the TDNN classifiers for varying $\tau_d$ and $n_h$ values are given in Table 2.2. We observe that TDNN achieves the best AUC performance as 87.2% with parameters $\tau_d = 16$ and $n_h = 20$. Yet, the other AUC performances with the number of hidden nodes 10 and 20 are all close to the best performance. The parameters for the final head-motion-based laughter detection system are fixed as $\tau_d = 8$ and $n_h = 20$, since they attain lower computational complexity and sustain 87.0% AUC performance. On the other hand, the SVM classifier attains 81.5% AUC performance with static head features. Hence, in our experiments with bagging and fusion, we keep using the TDNN classifier with the dynamic head motion features.

Table 2.2: The AUC performances (%) of TDNN classifiers with dynamic head features $\Delta f^H$ and with varying delay $\tau_d$ and number of hidden nodes $n_h$.

| | **Hidden Nodes** ($n_h$) | | |
|---|---|---|---|
| **Delay** ($\tau_d$) | **5** | **10** | **20** |
| **4** | 84.0 | 85.5 | 86.3 |
| **8** | 85.4 | 86.5 | **87.0** |
| **16** | 85.1 | 86.3 | 87.2 |

For laughter detection from facial motion, we consider both static and dynamic features, $f^P$, $F^P$, $\Delta f^P$ and $\Delta F^P$. We perform extensive experiments for the following three objectives: 1) to determine the best-performing classifier-feature combinations, 2) to set the best facial feature selection based on the discriminative analysis results presented in Section 2.3.4, and 3) to select the hyper-parameters of the TDNN classifiers.

For the first two objectives, we take two settings for the TDNN classifier, one with ($\tau_d = 4$, $n_h = 5$) and the other for a more complex structure with ($\tau_d = 4$, $n_h = 20$).

Then, we test all possible feature-classifier combinations (static vs dynamic and TDNN vs SVM) with varying number of features selected based on the mRMR discriminative analysis. In Figure 2.5, we plot the AUC performances of these combinations with varying feature dimensions.

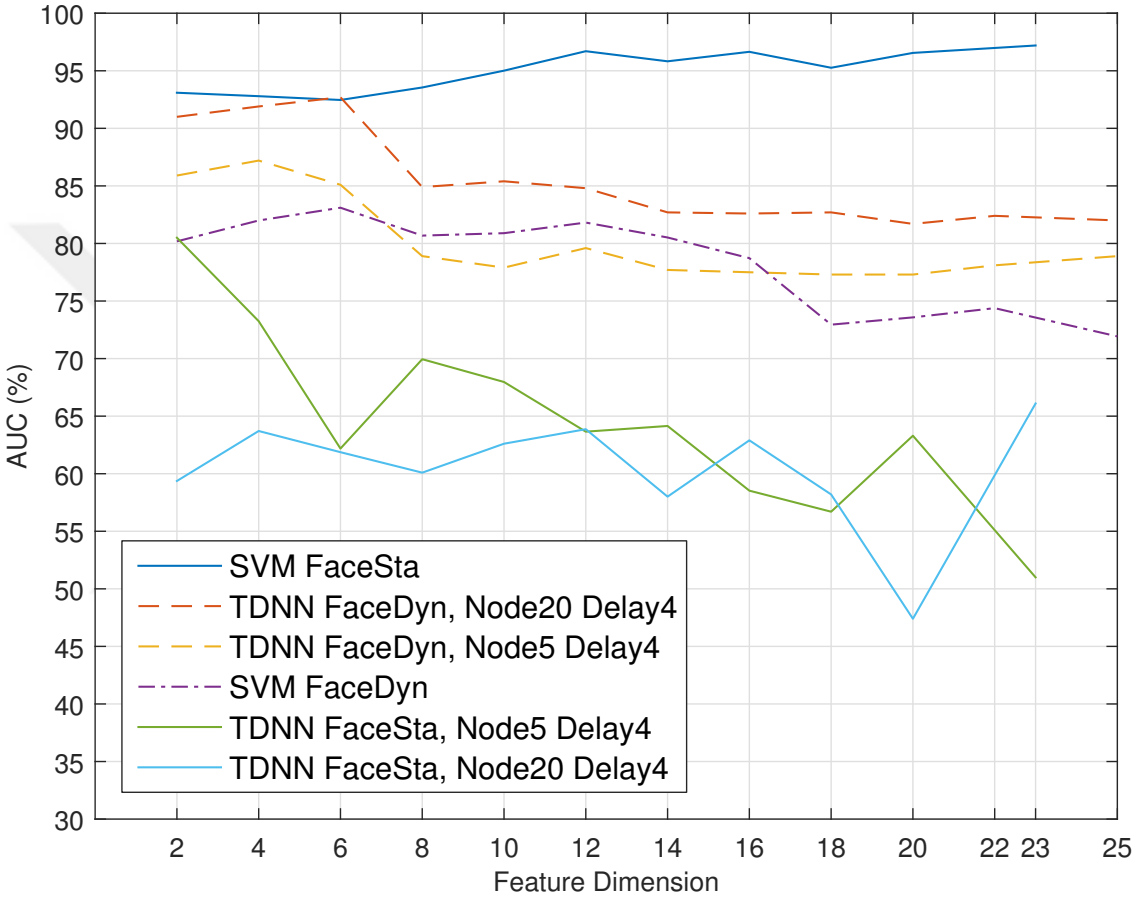



Figure 2.5: AUC performances of laughter detection using discriminative static and dynamic facial features at varying dimensions.

In Figure 2.5, we observe that for static facial features, the SVM classifier performs significantly better than the TDNN classifiers and attains its best AUC performance as 97.2% at feature dimension 23. Its performance saturates at this point and then starts to degrade slightly with 96.9%, 94.8% and 95.2% for feature dimensions 30, 40 and 50, respectively. Hence, in the upcoming experiments we employ the SVM classifier for the static facial feature $F^P$ with feature dimension 23.

As for the dynamic facial features, the TDNN classifiers outperform the SVM

classifier. When we consider the performance of the TDNN classifiers under two different hyper-parameter settings, we observe an overall peak at feature dimension 4. Hence, we set the TDNN classifier in the upcoming experiments with the dynamic facial features $\Delta f^P$ with feature dimension 4.

For the third objective, we evaluate the performance of the TDNN classifier for dynamic facial features over varying values of $\tau_d$ and $n_h$. Table 2.3 presents these performance evaluations. Since the AUC performances at $n_h = 20$ are higher compared to other $n_h$ settings and do not exhibit significant changes for varying values of $\tau_d$, we set the parameters as $\tau_d = 4$ and $n_h = 20$, which yield lower computational complexity due to a smaller delay parameter.

Table 2.3: The AUC performances (%) of TDNN classifiers with dynamic facial features $\Delta f^P$ at feature dimension 4 with varying delay $\tau_d$ and number of hidden nodes $n_h$.

| | **Hidden Nodes** ($n_h$) | | |
|---|---|---|---|
| **Delay** ($\tau_d$) | **5** | **10** | **20** |
| **4** | 87.2 | 90.9 | **92.2** |
| **8** | 86.2 | 91.1 | 92.4 |
| **16** | 86.1 | 89.4 | 92.1 |

Finally, we evaluate the performance of bagging for the best classifier-feature combinations, which are TDNN with dynamic head and facial features and SVM with static facial features. Figure 2.6 displays the ROC curves and AUC performances of these three classifier-feature combinations with and without bagging. Note that when bagging is incorporated, the performance of the TDNN classifiers with dynamic head and facial features improves, attaining respectively 88.3% and 92.5% AUC values. On the other hand, bagging does not help the SVM classifier with static facial features, which attains 97.2% AUC performance without bagging.

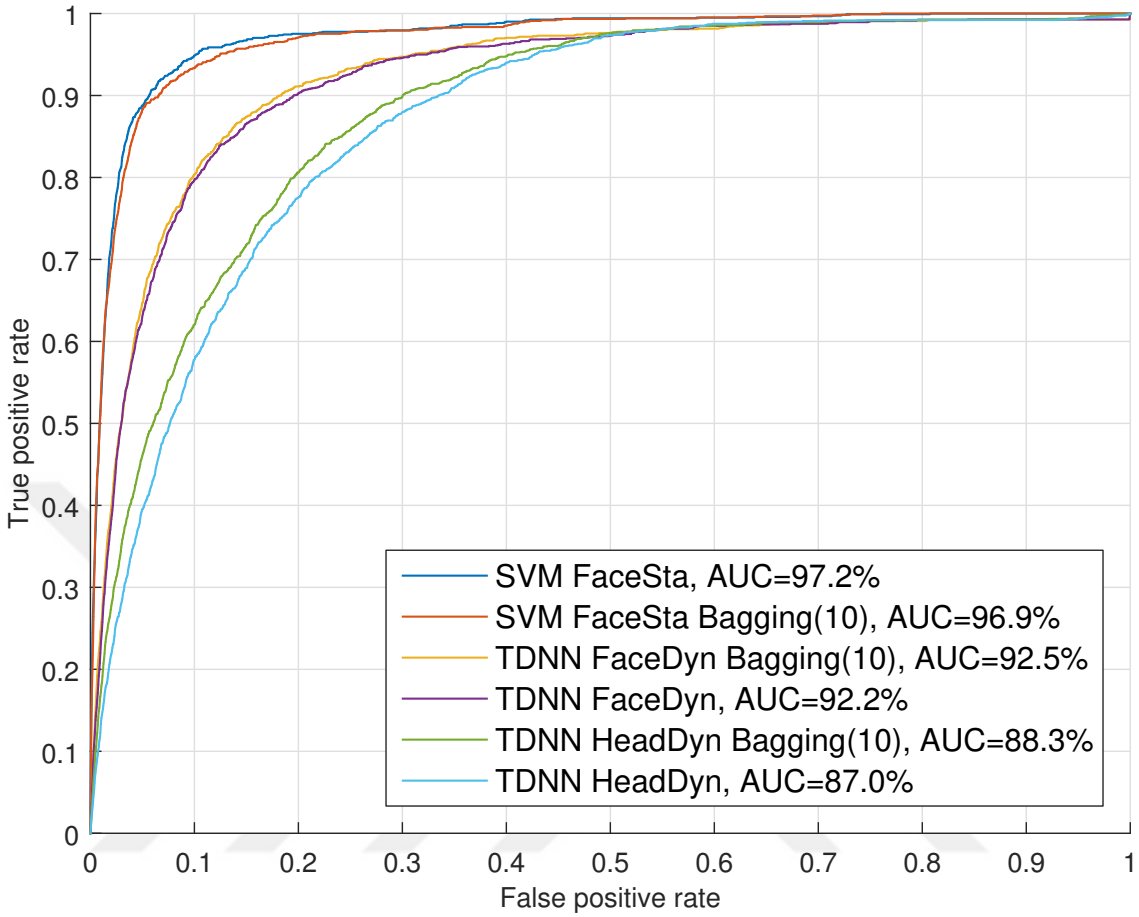



Figure 2.6: ROC curves and AUC performances of laughter detection from head and facial features using different classifier-feature combinations with and without bagging.

### *2.4.3 Fusion Results*

The best-performing classifier-feature combinations are integrated using the product rule defined in Section 2.3.5. Four classifier-feature combinations, SVM-$f^P$, TDNN-$\Delta f^P$, TDNN-$\Delta f^H$ and TDNN-$f^A$, are cumulatively populated, where all except SVM-$f^P$ are with 10-fold bagging. Figure 2.7 presents the ROC curves and AUC performances of three fusion schemes, which respectively have the best AUC performances for fusion of two, three and four classifiers. The best fusion scheme for two classifiers (Fusion2) is between SVM-$f^P$ (FaceSta) and TDNN-$f^A$ (Audio), which achieves 98.2% AUC performance. The best fusion scheme for three classi-

fiers (Fusion3) is between SVM-$f^P$ (FaceSta), TDNN-$\Delta f^P$ (FaceDyn) and TDNN-$f^A$ (Audio) with 98.3% AUC performance. Finally, the fusion of all four classifiers (Fusion4) attains 98.0% AUC performance.

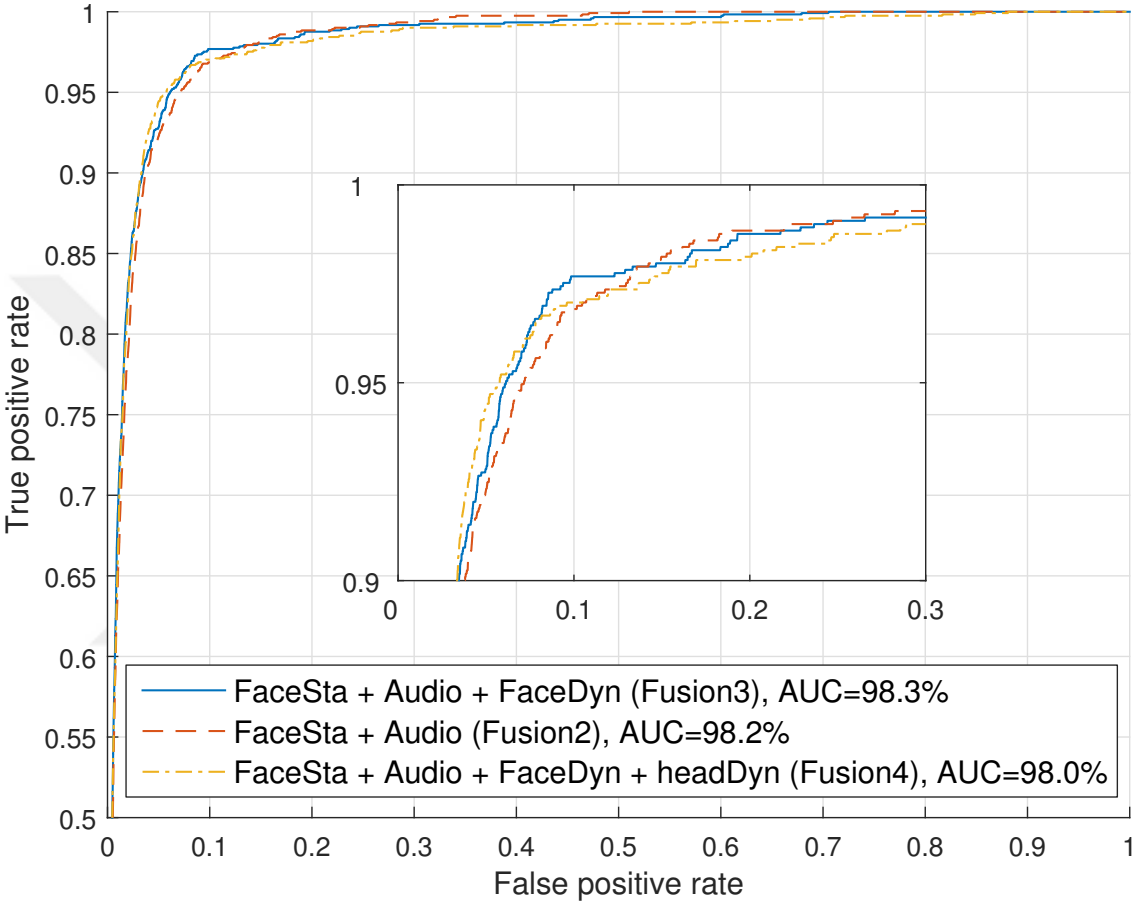



Figure 2.7: ROC curves and AUC performances of laughter detection with fusion of best classifier-feature combinations, face static (SVM-$f^P$), face dynamic (TDNN-$\Delta f^P$), head dynamic (TDNN-$\Delta f^H$) and audio (TDNN-$f^A$)

We further assess the performance of our audio-facial laughter detection schemes by using other common performance measures, such as recall, precision, and $F_1$ score. We set 2% false positive rate (FPR) as the anchor point on the ROC curve. At this anchor point, we first evaluate the recall performance under clean, cross-talk, and noisy laughter conditions in Table 2.4. We observe that the multimodal fusion of classifiers performs significantly better for all conditions. Static face features with SVM classifier perform reasonably well under cross-talk and noisy conditions, while

audio features with the TDNN classifiers suffer heavily in these cases. The Fusion3 scheme, which already has the best AUC performance, also yields the best recall performances except for the noisy condition. Note that dynamic head features with TDNN improve the multimodal fusion performance under noisy conditions, making the fusion of four classifiers, Fusion4, the most robust one against noise.

Table 2.4: Recall performances under all, clean, cross-talk and noisy laughter conditions for unimodal and multimodal classifiers at 2% FPR point

| | **All** | **Clean** | **Cross-talk** | **Noisy** |
|---|---|---|---|---|
| **Fusion3** | 83.2 | 80.4 | 87.6 | 86.5 |
| **Fusion4** | 82.7 | 80.2 | 86.6 | 88.5 |
| **Fusion2** | 79.0 | 77.1 | 82.3 | 82.7 |
| **FaceSta** | 73.1 | 67.9 | 80.0 | 80.8 |
| **Audio** | 40.9 | 50.1 | 29.2 | 34.8 |
| **FaceDyn** | 40.1 | 36.3 | 45.2 | 45.5 |
| **HeadDyn** | 28.0 | 27.4 | 28.9 | 48.5 |

Table 2.5 presents recall, precision, and $F_1$ score performances of the unimodal and multimodal classifiers at 2% FPR anchor point. Note that these performances are over all laughter conditions. As expected, the precision scores are lower than the recall scores due to data imbalance. Yet again, the best precision and $F_1$ score performances are obtained with multimodal fusion, where the best is the Fusion3 scheme.

### *2.4.4 Discussion*

We have reported the results of a detailed evaluation of our laughter detection scheme using various combinations of classifiers, modalities, and strategies. Below, we highlight some important observations and findings drawn from these experiments.

The discriminative analysis of laughter in Section 2.3.4 reveals that the facial laughter signal has a steady component, such as the contractions on the cheek region,

Table 2.5: Recall, precision and $F_1$ scores of laughter detection with unimodal and multimodal classifiers at 2% FPR point

| | **Recall(%)** | **Precision(%)** | $F_1$ **score(%)** |
|---|---|---|---|
| **Fusion3** | 83.2 | 26.5 | 40.2 |
| **Fusion4** | 82.7 | 26.4 | 40.0 |
| **Fusion2** | 79.0 | 25.5 | 38.6 |
| **FaceSta** | 73.1 | 24.1 | 36.2 |
| **Audio** | 40.9 | 15.8 | 22.8 |
| **FaceDyn** | 40.1 | 15.5 | 22.4 |
| **HeadDyn** | 28.0 | 11.4 | 16.2 |

which can be characterized by our positional (static) features, as well as a dynamic component, such as the abrupt and repetitive movements of head and mouth, that can be represented with our differential (dynamic) features. Our experiments also show that TDNNs can successfully capture temporal characteristics of laughter by relying on frame-level dynamic features, while SVMs can model the steady content via window-level summarization of static features. We observe that the SVM classifier with the static face features attains the best unimodal performance for laughter detection among all other classifier-modality combinations available.

According to AUC performance, the best classifier is Fusion3, which is the multimodal fusion of SVM with static face features, TDNN with dynamic face features, and TDNN with audio features. Note that although dynamic head features with TDNN attain 88.3% AUC performance, it does not improve the fusion of all classifiers, i.e., the Fusion4 classifier. However, it brings 2% improvement in the recall rate of noisy laughter segments, as observed in Table 2.4. Hence, head motion becomes valuable as a modality for laughter detection, especially in the presence of environmental acoustic noise.

In two aspects, our multimodal laughter detection scheme benefits from the discriminative facial feature analysis presented in Section 2.3.4. First, as observed

in Figure 2.5, the AUC performance drops more than 10% for the dynamic face features as the feature dimension grows further beyond a certain point. Second, feature dimension reduction helps to keep the training and testing complexities of the classifiers low, which allows the possibility of real-time implementations. In fact, the SVM and TDNN classifiers that we employ are well suited to real-time implementations with $O(M)$ time complexity, where $M$ is the size of the feature vector. The latency of the detection system in both cases is proportional to and actually a fraction of the window duration over which a decision is given. In the proposed framework, the worst mean computation times of the SVM and TDNN classifiers running in Matlab 2014b platform on a computer (Dell Latitude E5440) with Intel Core i7, 2.1 GHz CPU, are measured as 8.7 and 11.3 msec, respectively. Recently, we have utilized a real-time extension of the proposed laughter detection framework to analyze engagement in HCI [Türker et al., 2017a].

For the data imbalance problem, the bagging scheme has been investigated with both SVM and TDNN classifiers. We have observed that bagging brings much higher improvements for the TDNN classifiers. In general, TDNN architectures become harder to train as the number of parameters, the number of nodes, and delay taps increase. Although we have a fairly large laughter database, it does not allow us to employ larger TDNN structures. The balanced BLD database we use in training probably restricts the TDNN from better learning the non-laughter class. This could be the reason for the higher improvements that we obtain with bagging in the case of TDNN.

Existing work can be characterized as either recognition or detection frameworks. Recognition frameworks assume that the data has been pre-segmented into individual chunks, and the task reduces to classifying each chunk using standard classification algorithms. However, the data almost never comes in such a pre-segmented format. Hence, one needs to automatically segment the continuous data stream into smaller chunks and classify them. This is called detection. Detection is a far more challenging problem because it requires identifying segments in addition to recognition. Our method is a detection method, which sets it apart from most of the

existing work.

Three frameworks use only body movements [Griffin et al., 2013, Griffin et al., 2015, Niewiadomski et al., 2016] for performing recognition on pre-segmented laughter. Using various classification frameworks and different datasets, all three studies show that body movements convey valuable information for laughter recognition. The performance figures reported by these methods are comparable to those with audio and face modalities.

The first seven studies in the table [Ito et al., 2005, Petridis and Pantic, 2008, Reuderink et al., 2008, Escalera et al., 2009, Petridis et al., 2011, Scherer et al., 2012, Petridis et al., 2013a], use audio and face information in recognition and detection tasks. Two of them [Petridis et al., 2011, Petridis and Pantic, 2008] used pre-segmented laughter. In [Petridis and Pantic, 2008], although they perform a recognition task, unimodal and multimodal performances parallel our observations. That is, spectral features are the main source of high performance, while prosody features can provide additional enhancement up to some point. Also, the face modality performs better than audio, and the head movement information has the lowest performance. The work in [Petridis et al., 2011] describes a recognition system. However, the results are valuable since they come from a cross-database evaluation spanning four different datasets (AMI, SAL, DD and AVLC).

The next set of systems cover audio-facial laughter detection [Ito et al., 2005, Reuderink et al., 2008, Escalera et al., 2009, Scherer et al., 2012, Petridis et al., 2013a]. In [Reuderink et al., 2008], although the authors claim to present a detection method, the test results are reported on a relatively small set of pre-segmented data. The database used in this study is also unusually balanced (60 laughters and 120 non-laughters), atypical of the highly skewed distributions observed in naturalistic interaction. Furthermore, the evaluation in this work has been carried out by filtering away instances of smiles. This makes the dataset further biased by removing conceivable false positive candidates and thus potentially inflating performance. The best performance reported is 93% AUC-ROC for audio and face fusion.

In [Ito et al., 2005], the authors propose a laughter detection system and test

it on 3 clips of 4-8 minutes each. Unlike our work, the conversational data used in this study is not recorded in a face-to-face interaction scenario but through a video call between separate rooms. In addition, laughter instances constitute 10% of the whole data, which is quite high compared to ours (1.33%). This imbalance makes our task much more challenging.

Three threads of work [Escalera et al., 2009, Scherer et al., 2012, Petridis et al., 2013a] present proper detection algorithms on continuous data streams using relatively larger datasets. The method presented in [Escalera et al., 2009] uses only the mouth movements in the facial modality and there is almost no performance improvement in multimodal scheme over only audio modality. The work in [Scherer et al., 2012] contains 2 clips of 90 minutes dyadic conversation. The major limitation of this work is the lack of fine visual features in the data stream. A video of the face and the body was recorded simultaneously using a single omni-directional camera. The camera captures the entire scene, and all the participants in a single image, which makes it hard to capture fine level facial detail and reliable facial features. As such, the added benefits of the coarse visual information is unclear. Our work fills in the gap in this respect by demonstrating that the high resolution facial features based on tracked landmarks do improve the performance of laughter detection.

Finally, the work in [Petridis et al., 2013a] can be regarded as representing the 'state of the art' in audio-facial laughter detection. Here, the authors used the SEMAINE database with a fairly large test partition. In audio, they used MFCCs and in the visual modality they used FAPS information extracted from a facial point tracker. Recall, precision and $F_1$ scores are reported in a speaker dependent scheme. Also, they have a voice activity detector to get rid of silent regions in data. In agreement with our findings, they state that performance measures (recall, precision) suffer from class skewness (sparse positive class in natural interaction). They reported recall, precision, and $F_1$ scores of 41.9%, 10.3%, 16.4% respectively for the audio-facial scheme. If one seeks a comparison with our results, comparing $F_1$ scores would be meaningful. Our audio-facial $F_1$ score for Fusion2 is given in Table 2.5 as 38.6%, while they have reported audio-facial scheme $F_1$ score as 16.4%.

## *2.5 Summary*

We have introduced a novel audio-facial laughter detection system and evaluated its performance in naturalistic dyadic conversations. We have annotated the IEMOCAP database, which was originally designed to study expressive human interactions, for laughter events under cross-talk, noise and clean conditions. In this annotated database, we have investigated the utility of facial and head motion and audio for laughter detection using SVM and TDNN classifiers. For the face modality, we have used low dimensional and discriminative feature representations extracted using the mutual information-based mRMR feature selection method. Our experimental evaluations show that static motion features perform much better with the SVM classifier for the laughter detection task, whereas dynamic motion features as well as audio features perform much better with the TDNN classifier. One of the main findings of this work is that facial information as well as head motion is useful for laughter detection, especially under the presence of acoustic noise and cross-talks. Although the facial analysis in the presented system relies on the markers attached to skin, the marker positions are consistent with the MPEG-4 standard, and 3D tracking sensors such as Kinect can facilitate incorporation of these facial features into real-time laughter detection applications.

As future work, we think that the performance of our laughter detection system can further be improved using large scale training data, possibly by incorporating deep neural network architectures, such as recurrent neural networks.

# Chapter 3

# LAUGHTER RESPONSIVE ENGAGING ROBOTS

## *3.1 Introduction*

Engagement is a crucial component of user experience in human-computer interaction. Social agents need to build a bond with humans to retain their attention in conversations. Today, they still struggle to create and maintain the interest of individuals both in short and long time periods of interactions. Thus, understanding engagement and designing engaging agents is a step towards more naturalistic and sophisticated interactions.

With technological advancement, the robots' appearances have become more realistic, they possess more natural text-to-speech engines and can perform a plethora of complex tasks. These developments have contributed to abating the differences between human-robot and human-human interactions. However, they are still not sufficient for replacing human role in conversations with robots. Communication between two people consists of implicit and explicit channels that deliver essential signals to maintain the interaction as long as both parties desire. Agents should also be able to perceive, respond and make use of these signals. In this work, we focus on exploring the effect of laughter and smile as backchannels in human-robot interaction.

We design an interaction scenario involving two people and a back-projected robot head. In this scenario, the robot plays a quiz game with the participants. We conduct two sets of experiments, where the only difference is the mode of the robot - laughter responsive or laughter non-responsive. In the laughter responsive mode, the robot utilizes our real-time multimodal laughter detection module, to perceive laughter, and respond to it with laughter, or smile (if speaking). Whereas, in the

laughter non-responsive mode the robot does not respond to laughter by any means.

We evaluate the difference between the two kinds of interactions subjectively and objectively. For subjective evaluation, we use questionnaires to assess the participants' experiences. For objective evaluations, we measure the level of engagement of the participants' in both experiments by using the four connection events - directed gaze, mutual facial gaze, adjacency pair and backchannel, as described by Rich et al. [Rich et al., 2010].

Many of the existing studies on engagement have concentrated on the notion of engagement [Glas and Pelachaud, 2015, Peters et al., 2009], while the others have primarily focused on its measurement, detection and improvement in HCI. A recent survey summarizes the issues regarding engagement in human-agent interactions and presents an application on improving engagement on the GRETA / VIB platform [Clavel et al., 2016]. Being a comprehensive survey, this work emphasizes the importance of engagement in HCI and indicates the growing interest of researchers in the field.

Rich et al.'s work on engagement recognition is one of the pioneering studies in this area [Rich et al., 2010], where the authors propose an engagement model for collaborative interactions between human and computer. They conduct experiments on both human-human and human-robot interactions to have insight and evaluate their approach. Compared to their earlier work [Sidner et al., 2005b], they present a shorter list of dialog dynamics, which includes directed gaze, mutual facial gaze, adjacency pairs and backchannels. They refer to these four events as connection events (CE) between user and the robot, and use their timing statistics (min, mean, max of delays) to compare the engagement levels in two distinct scenarios, as well as an additional metric referred to as pace. Pace recapitulates the timing statistics, and it is inversely proportional to the mean time between connection events. The idea is that each CE refreshes the bond between human and robot, and increases the pace metric which is assumed to be proportional to the engagement level.

The backchannel, defined as a connection event, is an important aspect of engagement. As one of the social signals, it is a type of multimodal feedback, defined

by Yngve [Yngve, 1970] as "non-intrusive acoustic and visual signals provided by listener during speaker's turn". Humans, even unconsciously, respond to speaker using facial expressions, nodding, smiling back, using non-verbal vocalizations ('mm', 'uh-huh'), or verbal expressions ('yes', 'right'), which are all examples of backchannelling. There exist several studies which concentrate on backchannel timing prediction [Truong et al., 2010, Morency et al., 2010, Schroder et al., 2012, Meena et al., 2014], as well as various others addressing evaluation of backchannel timing such as [Inden et al., 2013, Gratch et al., 2007].

We specifically consider laughter as a backchannel signal in human-robot interaction. There exist other studies which integrate laughter in HCI and monitor its effect on the user. However, to the best of our knowledge, none of these studies evaluates the impact of a laughter responsive agent in terms of engagement. For example, Niewiadomski et al. experiment with a virtual agent in a simple interaction scenario with no verbal communication [Niewiadomski et al., 2013b], where subjects watch funny videos together with a laughter-aware virtual agent which mimics the subject's laughter. The humor experience of the subject is evaluated focusing on the quality of the synthesized laughter of the agent, and its aptness in timing. Another similar work is that of [Pecune et al., 2015], where the interaction scenario is also non-verbal. The subject and the virtual agent together listen to some funny music, and the agent mirrors the subject's laughter. The experience is then evaluated utilizing questionnaires. El Haddad et al. experiment with a virtual agent which can predict smile and laughter based on non-verbal expression observations from the speaker [El Haddad et al., 2016]. Nonetheless, their focus is on accuracy of the laughter prediction, and the naturalness of the synthesized laughter. They evaluate their system subjectively, using Mean Opinion Score (MOS).

## *3.2 System overview*

Our main objective is to analyze the role of laughter to engage users in human-robot interaction. Hence the robot should be able to perceive users' laughter and smiles in real-time and to respond back using these non-verbal expressions in its dialog flow.

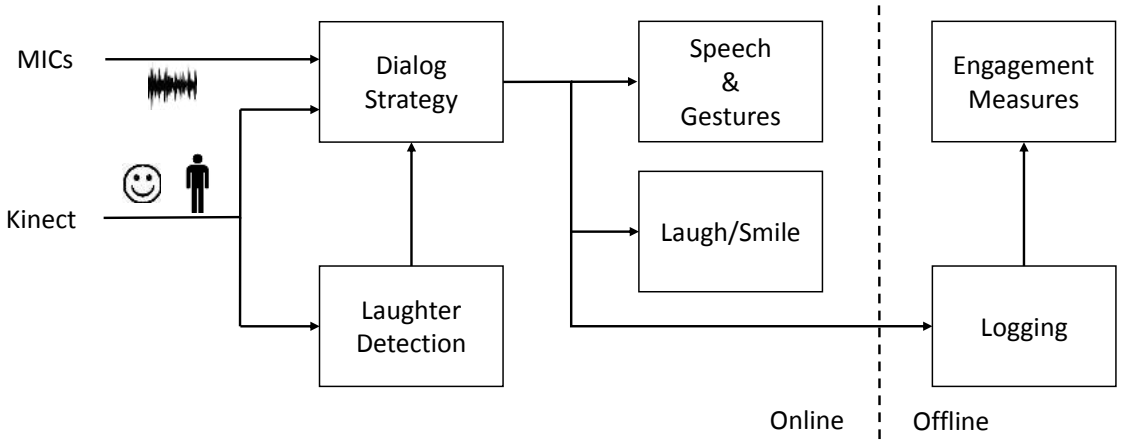



Figure 3.1: System overview

We hypothesize that such an ability of a robot will contribute to the engagement of the user during interactions.

Figure 3.1 shows the overview of the system. We have a dialog management block [Skantze and Al Moubayed, 2012], which takes user speech from microphones and positions from Kinect as inputs. It then creates a flow of dialog and gestures according to a rule-based strategy. The laughter detection module is involved with the dialog flow to trigger responses (laugh or smile) based on the detection results. All inputs and dialog flow components (produced gestures, speech etc.) are logged, and then processed so as to extract CEs and to compute engagement measures.

We employ a question-answer based scenario in which two subjects participate together and play a game of quiz [Skantze and Al Moubayed, 2012] with the robot. Basically, the robot starts with a short introduction. It asks participants' names and whether they know each other. The robot then proceeds with the quiz and poses some questions which are hard to guess but likely to draw attention from participants. An example question is "What color are sunsets on planet Mars?" and the given options are "green, blue, pink, orange". In order to finalize the quiz and the interaction, the robot keeps score of the correct answers of each participant, and declares the winner as the first to reach 3 points.

We exploit the fact that laughter mimicking by listener is the most natural

response to speaker's laughter. Therefore, during the interaction, when the robot is in its laughter responsive mode, it utilizes our laughter detection system for laughter detection. It thereby responds to laughter with laughs while listening, but with smiles while speaking (since it is hard to incorporate naturalistic laughter to speech). Whereas, in its laughter non-responsive mode, the robot does not perform laughter detection, and hence does not respond to laughs.

Figure 3.2 shows the experimental setup for the interaction scenario. The robot [Al Moubayed et al., 2012, Al Moubayed et al., 2013] (Furhat in this study) sits on one side of the round table. Participants are placed facing towards the robot. Kinect is on a tripod and able to see both participants' upper-body and face. Individual microphones are attached to participants' collars. Also, one video camera records the whole scene.

IrisTK platform [Skantze and Al Moubayed, 2012] is used with Furhat robot head, which provides functionalities, such as speech recognition and dialog management. On top of these modules, we build a real-time laughter detection module and engagement measurement methods, as we next explain in the sequel.

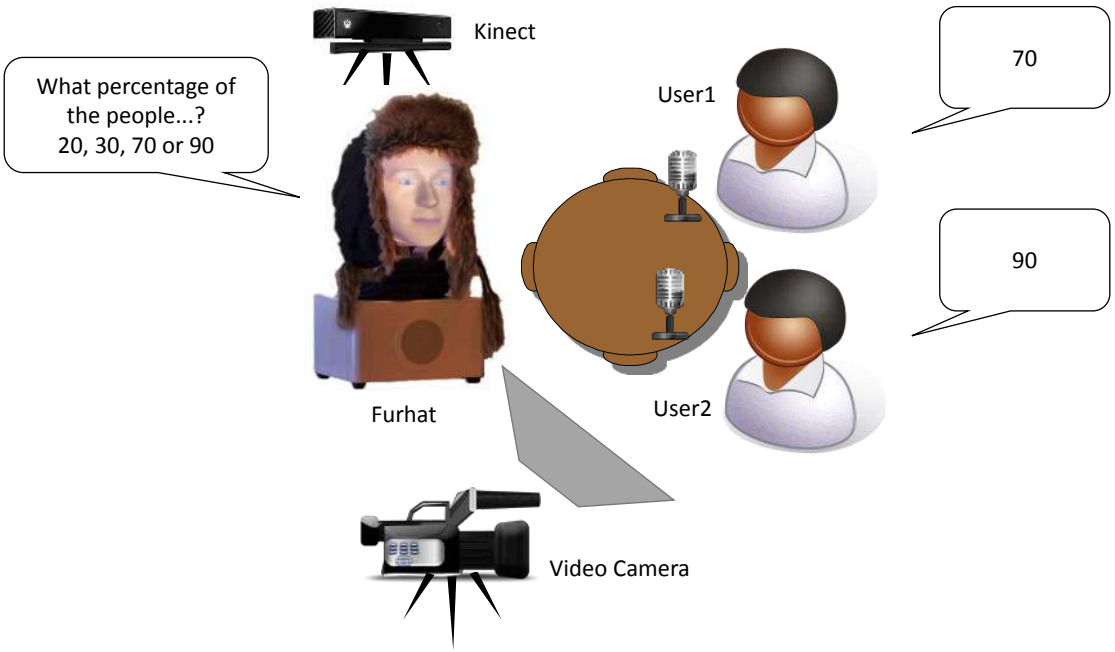



Figure 3.2: Experimental setup of the interaction scenario

## *3.3 Real-time laughter detection*

We have a multimodal scheme for laughter detection in naturalistic interactions [Turker et al., 2015, Türker et al., 2017b]. Audio and facial features are used to feed the detector. We have developed a method of detection on continuous audiovisual streams. It basically creates temporally sliding window on the stream and classifies with SVM whether the window instance involves laughter or not.

The detection method is trained over a human-robot interaction dataset [Devillers et al., 2015], which includes Kinect v2 data recordings. Kinect v2 provides whole body joints and facial landmark points along with high definition video and audio.

### *3.3.1 Audio and facial features*

We compute 12-dimensional MFCC features using a 25 msec sliding Hamming window at intervals of 10 msec. We also include the log-energy and the first order time derivatives into the feature vector. The resulting 26-dimensional dynamic feature forms the audio features.

Kinect can provide 1347 vertices of face model. We capture only 4 of them, corresponding to lip corners and mid points of lips, which roughly represent the lip shape. We keep these points in 3D coordinates to create a facial feature vector.

### *3.3.2 Summarization and classification*

Support vector machines (SVM) receive a temporal window of statistical summarization of the short-term features to perform binary classification for laughter and non-laughter classes. We also use probabilistic output of SVM classification for late fusion of modalities and setting different thresholds. The classification task is repeated for every 250 msec over overlapping temporal windows of length 750 msec.

### *3.3.3 Real-time implementation*

The implementation of laughter detection module is coded in C++ by using Kinectv2 libraries on Microsoft Windows. The process is designed to have two worker threads to handle data acquisition and feature extraction of modalities (audio and facial) and a master thread which fuses the outputs of the worker threads and produces the decision output.

In audio worker thread, the audio stream (16kHz sampling rate) is acquired in chunks of 256 samples. Hence, a two stage sliding window operation is performed in hierarchy. First, MFCC features are extracted over the audio buffer. The extracted MFCC features are then buffered and a sliding classification window is applied.

The video worker thread is similar to the audio thread but with a simpler feature extraction process. Kinect provides each visual frame (body, face, video etc.) with at most 30 fps. However, frames have their special time stamps rather than having fixed sampling period with (1/30) sec. The video thread grabs lip vertices each time a new frame arrives. Feature vectors are buffered where a sliding window runs over in order to have statistical summarization and SVM classification.

## *3.4 Engagement measurement*

We implement the methods proposed in [Rich et al., 2010] in order to measure engagement, which is applicable to face-to-face collaborative HCI scenarios.

Rich et al. have defined 4 types of 'connection events' (CEs) as engagement indicators:

- Directed gaze: Sharing the same location for both participants' gaze
- Mutual facial gaze: Face-to-face eye contact event
- Adjacency pair: The minimal overlap or gap between utterances (different speakers') during turn taking
- Backchannel: Backchanneling during other speaker's turn

Table 3.1: Interaction time statistics of the experiments

| | | **Interaction Time (sec)** | | |
|---|---|---|---|---|
| | **total #** | **min** | **max** | **mean** |
| **Laughter Resp.** | 10 | 138.0 | 296.6 | 222.1 |
| **Laughter Non-Resp.** | 10 | 126.2 | 515.1 | 267.8 |

We mostly follow the same methodology as [Rich et al., 2010] but with one small modification as we descibe in the following. In [Rich et al., 2010], there is only one participant interacting with the agent, and the directed gaze event is defined to happen when the agent and the participant look together at a nearby object related to the interaction. However, in our experiments, we have no objects of interest but an additional participant. Hence, when the robot, as a 'connection event initiator', changes its gaze direction from one participant to the other, this action initiates a 'mutual gaze' for one participant and a 'directed gaze' for the other.

In our experiments, CEs are extracted through the logged dialog components and sensory data (from Kinect). The extracted CEs are then used to calculate a summarizing engagement metric called 'mean time between connection events' (MTBCE). MTBCE measures the frequency of successful connection events. Basically, MTBCE in a given time interval $T$ is calculated by *$T$ / (# of CEs in $T$)*. As MTBCE is inversely proportional to engagement, similarly to [Rich et al., 2010], we use pace = 1/MTBCE to quantify the engagement between a participant and the robot. The pace measure is calculated over a range of different interaction durations such as the first 1 minute, the first 2 minutes and so on.

## *3.5 Experimental work and evaluation*

In the experiments, we used Furhat [Al Moubayed et al., 2012, Al Moubayed et al., 2013] as a conversational robot head. Furhat has the advantage of physical existence in the scene as well as having ability of efficient facial animation production.

At the beginning of each experiment, participants are briefly informed about

the experiment. They are told that they will simply play a quiz game with the robot. The operator explains the roles of the participants and the robot in the game without biasing them. Once participants are ready, they are left alone in an isolated experiment room.

In total, 20 experiments are performed in a randomly selected mode: laughter responsive or laughter non-responsive. A total of 10 experiments are conducted in each of the modes. Each experiment involves two people, therefore the engagement is evaluated over 40 subjects (28 male, 12 female, mean age: 25.9). The experiment ends when one of the participants reaches 3 points in the quiz (3 correct answers). The average time of an experiment is 4 minutes and 5 seconds. Table 3.1 indicates the statistics of interactions.

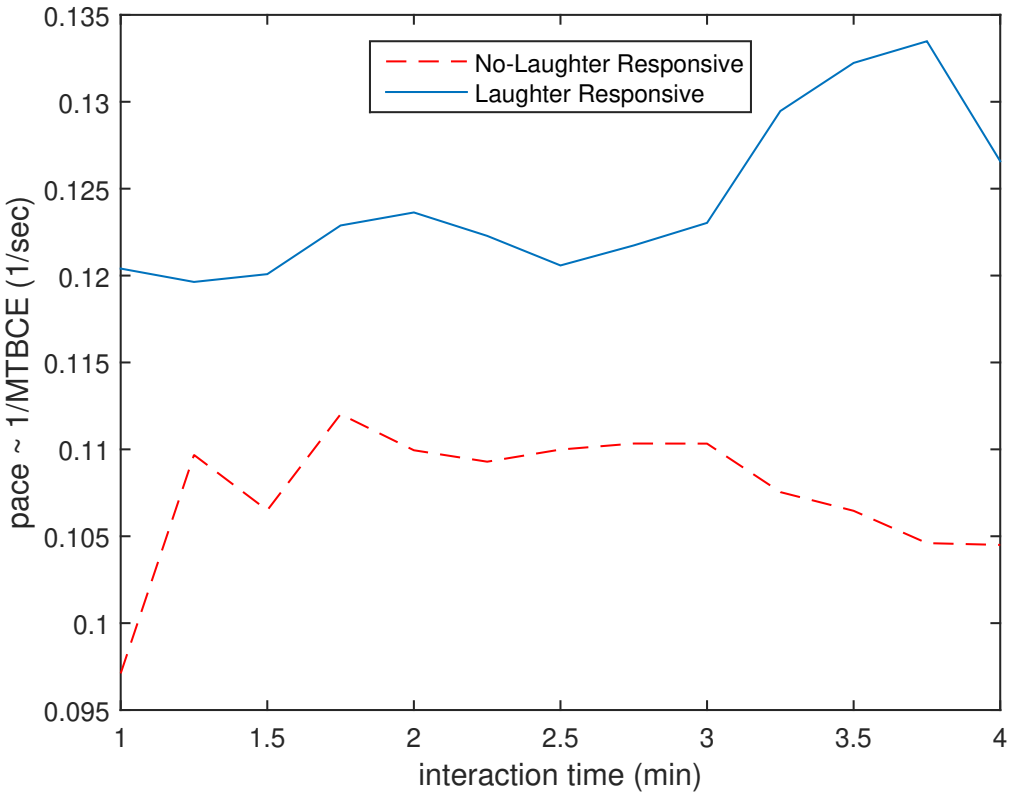



Figure 3.3: Average pace values for laughter-responsive (blue) and laughter non-responsive modes (red) over increasing interaction durations.

Table 3.2: Questionnaire items and mean scores over the laughter responsive and laughter non-responsive modes. Score scale: Strongly Agree (2), Agree (1), Undecided (0), Disagree (-1), Strongly Disagree (-2)

| | | **Responsive** | | **Non-Resp.** | |
|---|---|---|---|---|---|
| **#** | **Questionnaire Items** | **Mean** | **Std** | **Mean** | **Std** |
| **1** | I liked the interaction with the robot. | 1.50 | 0.60 | 1.40 | 0.60 |
| **2** | The interaction was entertaining. | 1.65 | 0.59 | 1.45 | 0.76 |
| **3** | I felt boredom at times during the interaction. | -1.35 | 0.59 | -1.00 | 1.03 |
| **4** | The robot was resp. to my emotional mood. | **0.65** | 0.75 | **-0.35** | 0.67 |
| **5** | The interaction felt natural. | 0.65 | 0.93 | 0.20 | 0.89 |

Figure 3.3 shows the average pace of the connection events over subjects in laughter responsive and laughter non-responsive mode of the robot. For example, the pace in the $n$-th minute is the average pace realized in the period from the beginning to the $n$-th minute of the interactions. We have calculated the pace for the first 4 minutes, as it is approximately the average duration of an interaction. We observe that the calculated pace values are significantly different between two modes for all interaction durations, indicating a considerable increase in engagement of a participant when interacting with a laughter responsive agent. We note that the pace samples belonging to the two modes exhibit statistically significant differences ($p << 6e-10$) when 2-sample t-test is applied. Also, the difference between the pace curves starts increasing after the 3-rd minute. This may be due to the observation that, after a warm-up period, participants tend to lose or increase their engagement according to the experiment mode.

For subjective evaluation, the subjects were required to fill a questionnaire after their interaction. Table 3.2 shows the five questions of the survey. We use a 5-point likert scale: Strongly Agree (2), Agree (1), Undecided (0), Disagree (-1), Strongly Disagree (-2) for each of the questions. To keep the subjects unbiased, even in the questionnaire, the fourth question implicitly asks if the users were aware of the robot's laughter response. The Mann Whitney test for the fourth question,

gives a statistically significant ($p = .0002$) difference between the laughter non-responsive and laughter responsive samples, which indicates that the users were aware of the laughter responsiveness of the robot during interaction. Question 5 also gives a statistically significant ($p = .05$) difference between the two samples, an evidence that laughter integration in HCI makes the interaction more naturalistic. The answers to other questions were not statistically different amongst the two samples. Nonetheless, this is expected because these questions are not intended to discern between the two modes of the robot, but rather to get feedback about the interaction scenario. Furthermore, since for most of the participants interacting with the robot was a first-time experience, they underwent the novelty effect. Simply put, even without laughter feedback from the robot, they enjoyed the interaction due to its novelty. Consequently, there is no statistically significant distinction for the first three enjoyment measuring questions.

Figure 3.4 plots the pace metric for each CE separately. All the CEs, except the mutual gaze (Figure 3.4b), yield higher pace curves for the robot's responsive mode. In the interactions, we observe the main reason behind the mutual gaze event loss: We discover that the majority of the participants, when amused, look whether the other participant is entertained, as well. In our scenario, this especially occurs when they are told their answer was wrong. In these occasions, the robot immediately shifts the attention from the current participant to require an answer from the other participant. Nonetheless, its initiation of mutual gaze event results in failure because the two participants are looking at each other.

Two strong tendencies of the responsive mode are observed in backchannel and directed gaze CEs. Figure 3.4d shows increase in the number of backchannel events, which are mostly laughter and smiles, in the second half of the interactions. Pace curves for the directed gaze events yield a decreasing trend for both modes, but responsive mode sustains higher pace values. This trend could be due to the participants' experience with Furhat. At the beginning of the experiment, participants are amazed when Furhat shifts his attention from one to the other participant by head and eye movement. Hence, participants tend to have successful directed gaze events

by looking at the other participant when Furhat does so. However, participants get acquainted with these attention shifts with time, which might be the cause of the decreasing trend of the pace for directed gaze events.

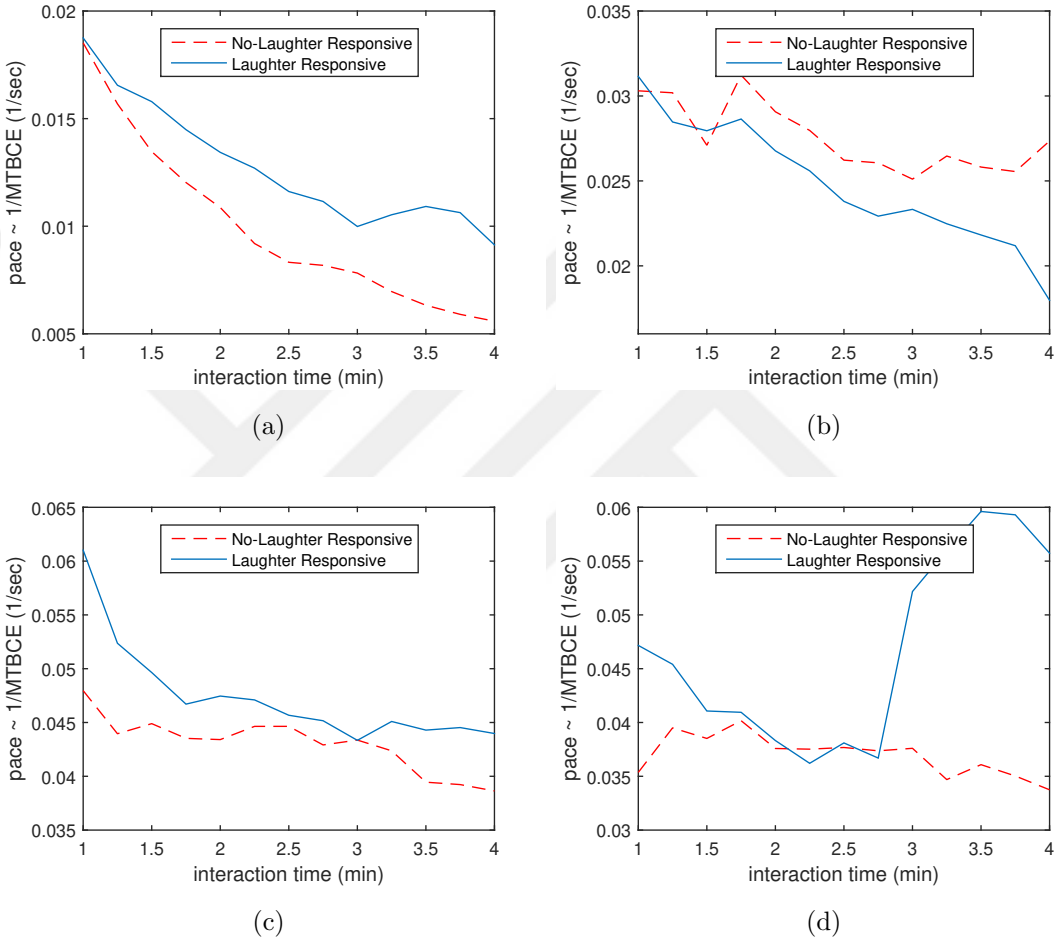



Figure 3.4: Average pace values using individual CEs: (a) Directed Gaze, (b) Mutual Gaze, (c) Adjacency Pair, (d) Backchannel

## *3.6 Summary*

In this work, we evaluated the effect of laughter in terms of engagement and user experience in human-robot interaction. In an interaction scenario with two people and a back-projected robot head, we experimented with two modes of the robot, laughter responsive and laughter non-responsive. In the laughter responsive mode,

the robot responds to subjects' laughter by laughter or smile, whereas in laughter non-responsive mode the robot does not respond to any laughter at all. We measure the engagement of the participants in two sets of experiments, objectively by utilizing the four connection events – directed gaze, mutual gaze, adjacency pair and backchannel. Our results indicate that the laughter responsiveness of the robot contributes to engagement of the participants. We also evaluate the user experience with a questionnaire, which likewise shows promising effects of laughter integration in an HCI system.

# Chapter 4

# MULTIMODAL PREDICTION OF HEAD-NOD AND TURN-TAKING EVENTS

## *4.1 Introduction*

Recent intelligent human-computer interaction (HCI) literature widely includes studies on language, dialogue management and speech recognition [Zue and Glass, 2000]. Targeting a more natural, flexible and generalizable HCI still sets challenging problems. Since early 2000s, inception of new research fields, such as investigation of non-verbal cues for human-human interaction, has enabled technologies for more humane (human-like) HCI systems [Clavel et al., 2016]. Robots and virtual agents are expected to understand what the user says, as well as to monitor the user's emotional and/or cognitive state and their audio/visual reactions (gestures, views, facial expressions, mimics etc.), to appropriately take more humane actions in the course of HCI. In this way, it is expected that the intermediaries will be more convincing and natural, and that they will keep the user engaged and enable them to interact more efficiently [D'Mello and Graesser, 2013, Schroder et al., 2012, Choi et al., 2012, Dybala et al., 2009, Bickmore et al., 2011, Bohus and Horvitz, 2014, Andre et al., 2004, Dybala et al., 2010].

In this study, we focus on head-nod and turn-taking events that are quite functional in human-human and human-robot interactions. As humans, we manage the conversational flow with smooth turn-takings and execute timely head-nods for emphasis or feedback. On the other hand, it is a challenging task to predict timely turn-taking or head-nod events to improve naturalness and user engagement in human-robot interactions. Furthermore, these two events help monitoring and sustaining the user engagement [Rich et al., 2010].

We construct a multi-modal framework for prediction of the head-nod and turn-taking events in dyadic conversations. The prediction task is basically a binary decision for a particular event which is likely to happen in the upcoming time instant. Hence, the problem can be defined as observing dyadic signals in time interval $[t-c, t]$ to make a prediction at time $t$ for the starting event at time $t + d$, where $c$ is the duration of the temporal window and $d$ is the time till the event start.

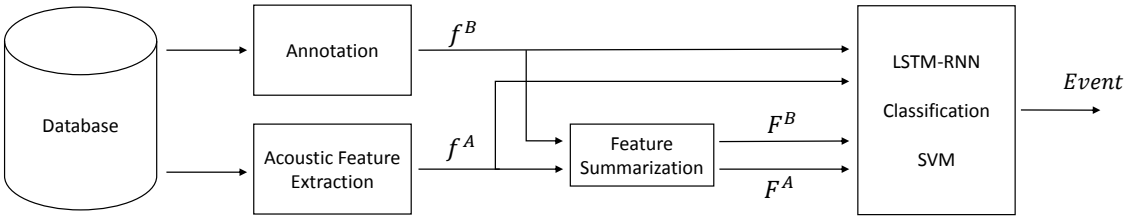



Figure 4.1: System overview

In the literature, head-nod recognition and detection have been studied extensively [Zhao and Allison, 2017, Kapoor and Picard, 2001, Chen et al., 2015]. However, head-nod timing prediction has been addressed in fewer number of studies.

In some of the works, head-nod is addressed under prediction of backchannels which are shortly defined as non-intrusive feedback expressions [Yngve, 1970]. The existing backchannel feedback mechanisms used in the human-computer interaction are often founded on rule-based approaches using simple statistical data [Schroder et al., 2012, Truong et al., 2010, Inden et al., 2013, Poppe et al., 2013, Poppe et al., 2014]. For instance, the relation between the changes in the sound perception of the speaker and the triggered backchannel signals of the listener has been examined in [Truong et al., 2010]. In this study, a backchannel signal is triggered when there are pauses with certain lengths, preceded by an increase or decrease in the sound pitch. In this estimation approach, the type of backchannel signal is not taken into consideration. In the SAL (Sensitive Artificial Listeners) system [Schroder et al., 2012], the engagement level and emotional response of the user is monitored, and events where backchannel feedback is necessary are observed. For example, when the user shakes their head or there is a change in their sound pitch, it is understood

that a feedback is necessary. According to the current mood state of the agent, through a predefined decision process a backchannel feedback such as smile or head nod/shake is synthesized.

There are very few studies in the literature aiming to learn backchannel by using a model based learning [Morency et al., 2010, Meena et al., 2014, Lee and Marsella, 2010]. For example, the head shake gesture as a backchannel is estimated from speech prosody, spoken words, and eye movements using Hidden Markov Models in [Morency et al., 2010]. Their estimation results are compared with reference data, but the proposed method is not implemented under any human-computer interaction scenario. Whereas in [Meena et al., 2014], only the verbal backchannel signals are predicted based on linguistic and prosodic cues using the naive Bayesian classifier. This classifier is tested on a verbal interface with a dialogue management mechanism, and it is reported to yield better results than an approach that randomly generates a backchannel signal. Again [Lee and Marsella, 2010], predict the head nod/shake by hidden Markov models in relation to verbal and emotional features of the interaction.

Turn-taking prediction is studied by many research groups since organization of the turns has been a problem in human-robot/agent interactions. Throughout the studies and observations in human-human interactions, prosodic changes [Kawahara et al., 2012, Skantze, 2017] and eye-gaze patterns [Novick et al., 1996, Kawahara et al., 2012] are shown that they are leading factors in turn-taking behaviours. Kawahara et al. [Kawahara et al., 2012], combine para-linguistic and non-verbal patterns to detect speaker change and to determine next speaker in poster sessions. They state that prosodic and gaze features are useful for speaker change detection, while backchannel of audience is also useful to determine next speaker. In a more recent study, Skantze [Skantze, 2017] proposed a continuous model solution for turn-taking problems. His work is solely based on audio features where visual channel communication is not considered. He used task-based interaction database in training while we use naturalistic social talk interactions which is more challenging and applicable in wild. Our work also contributes in prediction of head-nod which is a

social signal and supporting turn-taking predictions.

## *4.2 Methodology*

Our main objective in this study is to predict head-nod and turn-taking events before they take place in a dyadic conversation setup. We value these two problems for a more natural HCI system, in which a robot or virtual agent can predict probable head-nod or turn-taking opportunity by monitoring the interaction between human user and itself. In this purpose, we propose an event prediction framework, which can be trained using human-human dyadic conversational data. Figure 4.1 presents system overview of the proposed framework. There are two sets of features which are extracted from low-level acoustic signals and high-level non-verbal behavioral cues. We perform feature summarization for the SVM classifier and utilize temporal feature stream for the LSTM-RNN classifier. Each block of the framework and their input/output characterizations are given in the following subsections.

We define turn-taking event as the time instant that user ends her/his turn. Similarly, head-nod event is the time instant that head-nod action starts. Regardless of their type, we call them both '*event*' which occurs at given time $t$. Figure 4.2 shows event prediction structure over the interaction time-line of participants 1 and 2. Binary decision of the event occurrence is predicted by using features over the recent temporal window of length $c$ seconds.

We define two sets of features for the event prediction over the temporal window. Section 4.2.1 defines the first set as low-level acoustic features. The second set of features are defined on high-level non-verbal behavioral cues in Section 4.2.2. In the following subsections, aforementioned features and corresponding dimensions are provided for single participant. Eventual feature dimensions are given in Section 4.3, since minor changes apply for different event predictions: turn-taking and head-nod.

### *4.2.1 Acoustic Features*

Acoustic features are composed of both spectral and prosodic features. Spectral properties are described by mel-frequency cepstral coefficient (MFCC) representa-

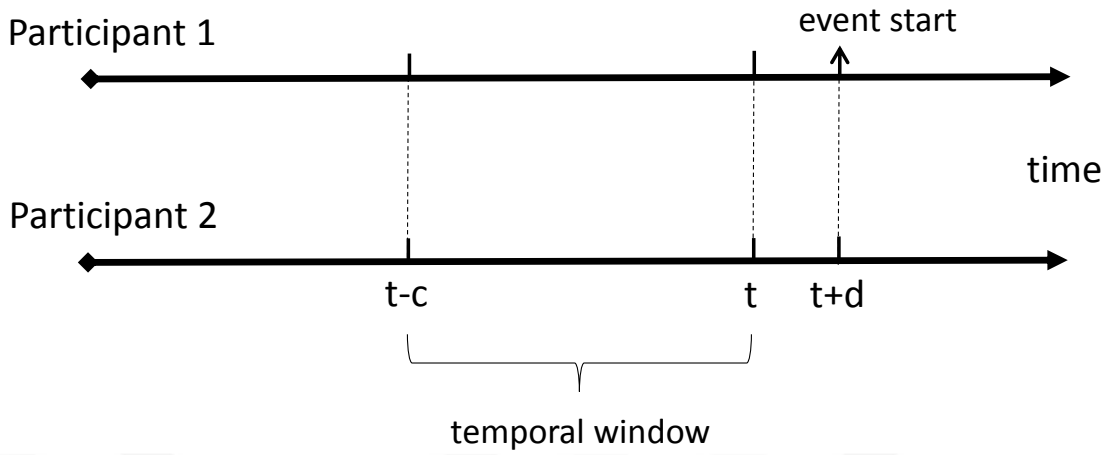



Figure 4.2: Event prediction time-line

tion which is the most commonly used spectral feature in speech and audio processing. We compute 12-dimensional MFCC features. Also, the log-energy coefficient is appended. Thus, the resulting 13-dimensional spectral feature vector is defined as $f^M$.

Prosody characteristics at the acoustic level carry important temporal and structural clues for audio portion just before the event occurrence. We choose to include speech intensity, pitch, and confidence-to-pitch into the prosody feature vector as in [Bozkurt et al., 2015, Bozkurt et al., 2016b]. Speech intensity is extracted as the logarithm of the average signal energy over the analysis window. Pitch is extracted using a well-known auto-correlation based method [de Cheveigne¥ and Kawahara, 2002]. Confidence-to-pitch provides an auto-correlation score for the fundamental frequency [Bozkurt et al., 2016b]. These three parameters and their first temporal derivatives form the 6-dimensional prosody feature vector, denoted by $f^P$.

Since extracted acoustic features are speaker and utterance dependent, we apply mean and variance normalization to both spectral and prosodic features. The mean and variance normalization of features is performed over the temporal window. Then the normalized spectral and prosodic features are concatenated and resulting 19-dimensional acoustic feature vector is obtained: $f^A = [f^M f^P]$. Note that acoustic features are extracted using a 40 msec sliding window at intervals of 25 msec. Thus,

feature frame rate is 40 Hz.

### *4.2.2 Non-verbal behavioral cues*

Non-verbal behavioral cues are produced by participants in the interactions, mostly in a unconscious way. These cues create a harmony between the participants, for instance one participant's head-nod may trigger the other's head-nod. Mirroring is defined as unconsciously copying others' non-verbal expressions in an interaction [Chartrand and Bargh, 1999]. Similarly, mirroring effect is observed in laugh/smile expressions [Finger, 2013]. On the other hand, gazing behaviours have important roles in social interactions. In [Ho et al., 2015], authors show eye-gaze convey essential cues for turn-taking. Also, turn-taking request can be expressed with relatively rapid, large and frequent head-nodding [Napier and Leeson, 2016].

Considering the meanings and their interactions of these non-verbal behavioral cues, we choose to include them into the feature set of our event prediction framework. The labels over the time-line is transformed into binary values with frame rate of 40 Hz. The frames get 1's where non-verbal behavioral event is active, the rest of the regions get 0's which means the event is not active. The frames are totally synchronous with acoustic feature frames. The resulting 3-dimensional non-verbal behavioral cues (head-nod, laugh/smile, gaze away) feature vector is denoted as $f^B$.

### *4.2.3 Feature Summarization*

We perform statistical feature summarization over the temporal window $[t-c, t]$ for the SVM classifier. For this purpose, we compute statistical quantities of the acoustic feature that comprise of 11 functionals, which are the mean, standard deviation, skewness, kurtosis, range, minimum, maximum, first quantile, third quantile, median quantile and inter-quantile range. This set of functionals were successfully used before by Metallinou et al. [Metallinou et al., 2013] for continuous emotion recognition from speech and body motion. Resulting summarized acoustic feature vector is denoted as $F^A$. The dimension of $F^A$ is 11 times the dimension of $f^A$.

Similarly, non-verbal behavioral cues features are reduced down to single vector.

Since these features are binary values, we choose to keep only one value which denotes the existence indication of the non-verbal behavioral cue in a given temporal window, $[t-c, t]$. For instance, if any portion of head-nod event appears within the temporal window, we summarize it as 1, otherwise as 0. Thus, summarized single vector has 3 dimensions and denoted as $F^B$.

### *4.2.4 Classifier*

We employ SVM and LSTM-RNN classifiers in the proposed event prediction framework. SVM is a binary classifier based on statistical learning theory, which maximizes the margin that separates samples from two classes [Cortes and Vapnik, 1995]. SVM projects data samples to different spaces through kernels that range from simple linear to radial basis function (RBF) [Chang and Lin, 2011]. We consider the summarized statistical features ($F^A$ and $F^B$) as inputs of the SVM classifier to discriminate occurance or absence of the 'event' at time $t + d$.

On the other hand, LSTM-RNN is an recurrent artificial neural network model. We use the LSTM-RNN classifier to model the temporal structure of the feature streams as it has been successfully used in many time-series data [Hochreiter and Schmidhuber, 1997b]. Since the model is sequential, frame based time-series features ($f^A$ and $f^B$) are used without summarization. In this work, we use only one LSTM layer. The further details of the classifier structure and its parameters are explained in Section 4.3.

## *4.3 Experiments and Results*

In Section 4.2, we describe the methodology of event prediction in dyadic interactions. Following that, we perform prediction of two events: turn-taking and head-nod. During the experiments, the person of interest (POI) is either Participant 1 or Participant 2. In other words, we have a single classification structure which takes one participant on focus and make prediction for that participant. For instance, each annotated head-nod belongs to one participant position: either 1 or 2. Thus, it is known which participant is POI for prediction of each head-nod. Then, we have

2 sources of features: POI and the other (OTH). Naturally, acoustic features and non-verbal behavioral cues features are available for both of them.

In turn-taking prediction, we use POI($f^A$, $f^B$) and OTH($f^B$). Since, we know only POI has speech in temporal window. On the other hand, in head-nod prediction we use all available features: POI($f^A$, $f^B$) and OTH($f^A$, $f^B$). Note that feature dimension orders are preserved both in between POI, OTH and $f^A$, $f^B$.

### *4.3.1 Dataset and Annotations*

In order to carry out experimental work, IEMOCAP database [Busso et al., 2008] is used. IEMOCAP consists of naturalistic human-human dyadic conversations with rich affective contents. Five dyads (sessions), interacts under scripted and spontaneous scenarios, resulting with 8 hours of data.

Annotations of non-verbal behavioural cues are carried out on IEMOCAP database. Laugh/smile annotation performed in our previous study [Turker et al., 2017]. Head-nod and gaze away cues are annotated by two human subjects. Then, the annotations are checked and if necessary corrected by a third subject.

In our annotation scheme, head-nod is defined as vertical swing of head for a reasonable of time with a conscious or unconscious mission of carrying a message to other interlocutor. This definition helped eliminating highly speaker dependent head-nod behaviours. On the other hand, gaze away is defined as the time portions that the gaze has a fixation out of the other interlocutor. Rapid eye movements (saccades) are not annotated as gaze away.

Table 4.1 reports basic annotation statistics of the three non-verbal behavioral cues and turn segments that are longer than 5 seconds. Turn annotations comes with IEMOCAP database as described in [Busso et al., 2008].

### *4.3.2 Training*

We set the temporal window size $c = 2$ sec for turn-taking and $c = 3$ sec for head-nod event prediction. Then, we create a class balanced dataset for each events. For head-nod events, we obtain balanced dataset by randomly picking 1648 negative class

Table 4.1: Annotation statistics of the three non-verbal behavioral cues and turn segments that are longer than 5 seconds

| | **# of events** | **Duration (sec)** | |
|---|---|---|---|
| | | **Mean** | **Std** |
| **Head-Nod** | 1648 | 1.25 | 0.67 |
| **Laugh/smile** | 1244 | 0.96 | 0.91 |
| **Gaze Away** | 5147 | 4.26 | 5.14 |
| **Turns** | 3132 | 8.04 | 3.02 |

samples from no head-nod regions. For turn-taking events, for each turn region, we extract one positive $[t-2, t]$ and one negative sample $[t-4, t-2]$. We keep only turns longer than 5 sec since we experiment with varying $d = \{0, 0.1, 0.2, 0.4, 0.6, 0.8, 1\}$. We attain balanced dataset with 6264 samples. In all experiments, these balanced datasets are used in one-session-out (5-fold) test fashion. Thus, all results are speaker independent.

RBF kernel is used in SVM training and hyper-parameters are optimized in the first fold with grid search over cross-validation scheme. Optimized parameters are kept fixed for the rest folds.

In training of LSTM-RNN, we used single layer of LSTM with 50 hidden nodes. During the 5-fold test, in each fold we split 1 session for validation and train with 3 sessions. Early-stopping is provided by the best accuracy performance on validation set.

### *4.3.3 Results*

In this section, we report our experimental evaluations, which are obtained over the balanced datasets created in Section 4.3.2. Since, experiments are over class balanced set, widely used performance metrics, accuracy (Acc), precision (Pre), recall (Rec), $F_1 - score$ $(F_1)$, are utilized in the evaluations. Note that, positive class is happening of the event: head-nodding of POI, turn-ending of POI (turn-

taking of OTH).

Table 4.2: Turn-taking prediction performances with the SVM for varying $d$ in seconds

| $d$ | **0** | **0.1** | **0.2** | **0.4** | **0.6** | **0.8** | **1** |
|---|---|---|---|---|---|---|---|
| **Acc** | 75,69 | 73,65 | 71,57 | 65,72 | 58,96 | 57,27 | 54,92 |
| **Pre** | 74,98 | 72,79 | 70,71 | 64,92 | 58,13 | 56,36 | 54,38 |
| **Rec** | 77,10 | 75,56 | 73,65 | 68,40 | 64,11 | 64,43 | 61,11 |
| $F_1$ | 76,03 | 74,15 | 72,15 | 66,61 | 60,97 | 60,13 | 57,55 |

Turn-taking prediction performances with SVM classifier are given in Table 4.2. The best performances are observed with the minimum delay till the event, $d = 0$, which is expected by the nature of the problem. We expect the largest differentiation between positive and negative class samples when $d = 0$ and thus better performance than the others, $d > 0$.

Considering human-robot interaction, giving 200 msec advance to the robot to take the turn could be very beneficial even though prediction performances degrade slightly at $d = 0.2$ sec. Another promising observation is having close precision and recall performances since both are important for the task. High precision means less false positives and thus less intervention while user's turn continues. High recall means less false negatives and thus less lagging to take the turn.

Table 4.3: Turn-taking prediction performances with the LSTM-RNN for varying $d$ in seconds

| $d$ | **0** | **0.1** | **0.2** | **0.4** | **0.6** | **0.8** | **1** |
|---|---|---|---|---|---|---|---|
| **Acc** | 83,62 | 81,91 | 79,89 | 72,94 | 64,45 | 61,14 | 58,42 |
| **Pre** | 80,53 | 78,21 | 77,51 | 71,59 | 63,47 | 61,64 | 57,98 |
| **Rec** | 88,68 | 88,46 | 84,23 | 76,07 | 68,07 | 58,98 | 61,14 |
| $F_1$ | 84,41 | 83,02 | 80,73 | 73,76 | 65,69 | 60,28 | 59,52 |

Table 4.3 presents turn-taking prediction performances using LSTM-RNN. We observe remarkable performance improvement with LSTM-RNN compared to SVM

for small $d$. As $d$ increases, performance difference is getting smaller, which is quite legitimate since negative and positive class samples getting so similar and makes hard to capture pattern differences for both of the classifiers. Another observation is that LSTM-RNN produces systems with better recall performances. This can be useful for a robot, which displays more talking with more generous interventions but also timely turn-takings.

Table 4.4: Head-nod prediction performances with SVM and LSTM-RNN

| | **SVM** | | **LSTM-RNN** | |
|---|---|---|---|---|
| | $F^A, F^B$ | $F^A, F^B, F^{SA}$ | $f^A, f^B$ | $f^A, f^B, f^{SA}$ |
| **Acc** | 59,46 | 62,49 | 59,01 | 63,37 |
| **Pre** | 60,10 | 63,53 | 61,75 | 64,21 |
| **Rec** | 56,13 | 58,56 | 47,21 | 60,32 |
| $F_1$ | 58,05 | 60,94 | 53,51 | 62,20 |

In head-nod prediction task, we consider speech activity (turn states) to help prediction of head-nods, as turn-taking prediction already owns head-nod features. Thus we extract speech activity feature $f^{SA}$, which is binary valued stream obtained from turn annotations as defined for the other non-verbal behavioral cues features. Similarly, summarized speech activity feature $F^{SA}$ is just defined as a binary value to indicate any temporal change in $f^{SA}$. The $F^{SA}$ indicates if the participant started talking or stopped talking.

In Table 4.4, head-nod prediction performances are given with and without the speech activity features. We observe that speech activity features improve performances for both classifiers, but LSTM-RNN favors more than SVM. Although unimodal head-nod prediction performances of acoustic features and social cues are poor and close to random, the multimodal performances are promising. Since precision is higher, it indicates less false positives and thus less awkward timely head-nods. False negatives do not create much trouble, since passing head-nod opportunity as a robot is not a big problem. Actually, even humans have much variety

on head-nodding behaviour. One can just pass head-nodding if she/he experienced very same moment that she/he head-nodded before.

We perform decision fusion of the best performing SVM and LSTM classifiers by weighted sum of the classifier confidence scores as, $\alpha * SVM(F^A, F^B, F^{SA}) + (1 - \alpha) * LSTM(f^A, f^B, f^{SA})$, where $\alpha$ is the weight of the SVM confidence score. Table 4.5 presents head-nod prediction performances of the decision fusion for three values of $\alpha$. We observed that the decision fusion with $\alpha = 0.6$ produced the highest $F_1 - scores$ for head-nod prediction task.

Table 4.5: Decision fusion of the best performing SVM and LSTM classifiers: SVM($F^A, F^B, F^{SA}$) and LSTM($f^A, f^B, f^{SA}$)

| $\alpha$ | **0.5** | **0.6** | **0.7** |
|---|---|---|---|
| **Acc** | 64,28 | 65,28 | 65,34 |
| **Pre** | 65,56 | 66,80 | 67,17 |
| **Rec** | 60,07 | 60,68 | 59,95 |
| $F_1$ | 62,70 | 63,59 | 63,35 |

## *4.4 Summary*

In this study, we present a generalized framework for event prediction in dyadic interactions, such as human-human and human-robot. We report turn-taking and head-nod prediction performances over human-human conversational data. We showed that turn-taking prediction can be achieved with relatively better performance, even for predictions hundreds of milliseconds ahead. Head-nod prediction experiments showed that speech activity features have potential to improve the performance and fusion of classifiers achieves the best overall performance.

The proposed framework has a potential use for more humane human-robot interactions. Smooth turn-takings and producing timely head-nods are expected to make robots more natural and engaging.

# Chapter 5

# LAUGHTER PREDICTION USING MULTI-MODAL TRANSFORMERS

## *5.1 Introduction*

Non-verbal cues, particularly smiles and laughs, are intrinsic components of human communication. They facilitate various communicative functions and contribute to the complex social signals that enrich human interaction (Ruch, 2008; Ekman et al., 1987). The emulation of these cues in robotic systems is pivotal to establishing natural and engaging human-robot interaction (HRI). With the advent of machine learning techniques, the prediction and generation of these non-verbal cues have become feasible, but still present significant challenges (Kim et al., 2013).

The primary objective of this research is to develop a robust machine learning model capable of predicting upcoming candidate smiles as backchannels in dyadic conversations. By utilizing the mimicry dyadic conversation database, a comprehensive approach is adopted, employing Long Short-Term Memory (LSTM) networks and cross attention Transformer models. This multi-modal analysis contrasts the effects of relatively long-term windows for speech against short-term windows for acoustic cues.

The key contributions of this work are as follows:

1. The Development of a Transformer-Based Model: Leveraging the potential of cross attention mechanisms, this research introduces a novel application of Transformer models in predicting smiles in HRI.

2. Comparative Analysis: The study provides a thorough comparison between LSTMs and Transformer models, highlighting the substantial improvements

achieved by the latter.

3. Utilization of Multi-modal Features: By integrating both speech and acoustic cues within distinct temporal windows, the model ensures a nuanced understanding of non-verbal communication patterns.

This research aims to provide a comprehensive study on the prediction of smiles as backchannels in dyadic interactions, with significant contributions to the field of HRI. The methodological approach and insights gleaned from this research offer promising avenues to improve the naturalness and engagement of robotic systems.

## *5.2 Related Work*

Smiles and laughs serve as complex social signals that embody various functions in human interaction, playing a crucial role in communication and expressing emotions such as joy, humor, and social bonding. Understanding the psychological roles and interpretations of these non-verbal cues is essential for effective human-robot interaction (HRI). Ruch [Ruch, 2008] examined the psychological roles of laughter, linking it to the expression of humor and joy [Dunbar, 2012]. Laughter has been well-studied in scientific literature, including its function in human conversation and interaction [Inoue et al., 2022]. It is known to trigger positive emotions and enhance social bonding. However, cultural variations exist in the appropriateness and interpretations of smiles and laughter. Elfenbein et al. [Elfenbein, 2007] investigated these variations, revealing differences in the interpretations of smiles and laughter across cultures [Dunbar, 2012]. Non-verbal expressions, such as smiles, can carry different meanings and implications in different social contexts [Dunbar, 2012].

Ekman et al. [Ekman et al., 1987] categorized smiles into different types based on emotional context and genuine expression, leading to various interpretations of smiles in communication [Dunbar, 2012]. This categorization provides insights into the nuanced meanings conveyed by different types of smiles, which is crucial for effective communication in HRI. The use of machine learning algorithms for predicting non-verbal cues in HRI has been extensively studied. Kim et al. [Schuller

et al., 2013] explored decision trees, support vector machines (SVMs), and neural networks for gesture prediction.

Niewiadomski et al. [Niewiadomski et al., 2013a] utilized Multimodal Hidden Markov Models to generate smiles and laughter in virtual characters. These studies showcase the growing interest in incorporating laughter and smiles into interactive systems, aiming to enhance the naturalness and effectiveness of human-robot interactions. The acoustic properties of speech, such as pitch and intensity variations, have been employed to detect and predict laughter in conversational speech [Szameitat et al., 2022]. Facial expressions and body language are also essential visual cues for laughter prediction [Hofmann et al., 2017].

Recent studies have emphasized the role of contextual information in generating smiles and laughter. Ishi et al. considered conversational context in generating smiles, highlighting the importance of situational awareness in prediction models [Ishi et al., 2019]. Timing and temporal dependencies are also crucial factors in natural human-robot interaction. Morency et al. [Morency et al., 2008] applied Hidden Markov Models to capture sequential patterns. Engagement is a metric that has been advocated for assessing interaction quality in HRI. Sidner et al. [Sidner et al., 2005a] proposed engagement as a metric, and Dautenhahn [Dautenhahn, 2007] correlated engagement with perceived empathy and naturalness of robotic behavior. User studies have played a pivotal role in understanding human perception of smiles and laughs in HRI. Breazeal [Breazeal, 2003] conducted empirical studies on the effectiveness of facial expressions in robotic characters, while Bickmore and Picard [Bickmore and Picard, 2005] evaluated user satisfaction with virtual agents exhibiting various non-verbal behaviors.

Real-time processing and personalization are significant challenges in smile and laugh prediction. Paiva et al. [Paiva et al., 2017] stressed the importance of personalization in adapting models to individual users. The literature provides a multifaceted view of smiles and laughs as communicative signals in human interaction and HRI. It covers diverse modeling techniques, feature selections, evaluation methods, and emerging challenges. The understanding of the psychological roles, cultural

variations, and interpretations of smiles and laughter is crucial for effective communication in HRI. The development of accurate prediction models, considering acoustic, visual, contextual, and temporal cues, can significantly enhance the effectiveness of HRI systems.

## *5.3 Database*

In our analysis and experimental evaluations, we have used the Mimicry Dyadic Conversation Database [Bilakhia et al., 2015]. The Mimicry database was created to meet human-human interaction recordings of everyday situations, with a particular focus on the analysis of mimicry in human-human interactions. The aim of the database is to provide a collection of recordings that demonstrate a considerable amount of similarity and/or synchronization between the participants.

The recordings were conducted in a laboratory setting with 15 cameras and 3 microphones to ensure the best possible conditions for analyzing the behavior. All sensory data was precisely synchronized (with a margin of error of less than 10 nanoseconds) through hardware triggering. Two experiments were recorded: a political discussion and a role-playing game. In total, 54 recordings were made, 34 of the discussion and 20 of the game. Forty participants were recruited, 28 male and 12 female, aged between 18 and 40. After the experiments, all participants reported their feelings.

### *5.3.1 Laughter Annotation*

An extensive annotation effort was conducted by a team comprised of three individuals with expertise in audio-video annotation. Manual annotation of prominent smiles and laughter was performed to ensure a high degree of accuracy in the ground truth.

Annotations were executed using the ELAN Annotation Tool [Wittenburg et al., 2006], with each annotator responsible for approximately 350 minutes of video footage. The annotation process involved iterative cycles of watching, pausing, re-watching, and demarcating boundaries, followed by fine-tuning through multiple

replays. Due to the meticulous nature of this process, a real-time factor of 2 was incurred, meaning that approximately 700 minutes were required to annotate each 350-minute segment of video. The third member of the team, then watch the all annotated clips and double check the final outcome of the work.

In total, 706 laughter and 149 smile instances are annotated. Average duration of the segments is 1338ms and standard deviation is 1077ms.

### *5.3.2 Transcription*

All audio files within the database were transcribed utilizing an Automatic Speech Recognition (ASR) system, specifically provided by Google. Despite the dual-channel recording facilitated by two separate microphones, the presence of cross-talk was observed across both channels, a consequence of the inevitable proximity of the microphones. Consequently, speaker diarization could not be effectively implemented for the transcription process.

### *5.3.3 Large Conversational Transcription Data: OpenSubtitles*

To assemble a substantial corpus of conversational text data, we employed the OpenSubtitles dataset [Lison and Tiedemann, 2016], primarily comprised of subtitles from films and television series. Particularly useful were the subtitles designed for individuals with hearing impairments, which often contain textual annotations indicative of laughter, such as '[laughs]' and '[giggles]'. It should be noted that these annotations are not derived from a standardized lexicon but rather vary depending on the author or content provider. Therefore, a comprehensive examination was conducted to identify laughter-associated annotations using regular expressions. In total, 532,044 instances featuring laugh-positive labels were extracted.

## ***5.4 Proposed Method***

The proposed method utilizes a cross-attentional multi-modal Transformers model to predict smiles/laughs as backchannels in dyadic conversations. The model uses

speech and acoustic cues processed through separate temporal windows to create a nuanced representation of conversational dynamics.

1. Long-Term Window for Speech: Speech cues are processed through a longer temporal window, capturing the conversational context.

2. Short-Term Window for Acoustic Cues: Acoustic cues are processed through a shorter window, reflecting immediate reactions and emotional states.

### *5.4.1 Acoustic Features*

In the context of acoustic feature extraction, two primary approaches are considered: one involves leveraging the raw waveform, allowing the neural network layers to intrinsically capture temporal features, and the other entails the extraction of spectral and prosodic features, which are subsequently fed into the model. These alternative methodologies are employed under various experimental conditions, the specifics of which will be elaborated upon in the section 5.5.

The Mel-Frequency Cepstral Coefficient (MFCC) serves as the preeminent spectral feature in the realm of speech and audio processing. In this work, 12-dimensional MFCC features are computed utilizing a 25-millisecond sliding Hamming window with a 10-millisecond overlap. Additional components, such as log energy and first-order time derivatives, are incorporated to construct a 26-dimensional feature vector representing the spectral content.

At the level of prosodic characteristics, which encapsulate elements like intonation, rhythm, and intensity patterns, several key features are selected for inclusion in the feature vector. Specifically, speech intensity is determined through the logarithm of the average signal energy over the analysis window. The fundamental frequency, or pitch, is assessed using the YIN algorithm, a widely acknowledged auto-correlation-based estimator. An auto-correlation score is also computed to quantify confidence-to-pitch.

Given that prosodic features are both speaker-dependent and utterance-specific, mean and variance normalization are applied to mitigate these variabilities. This

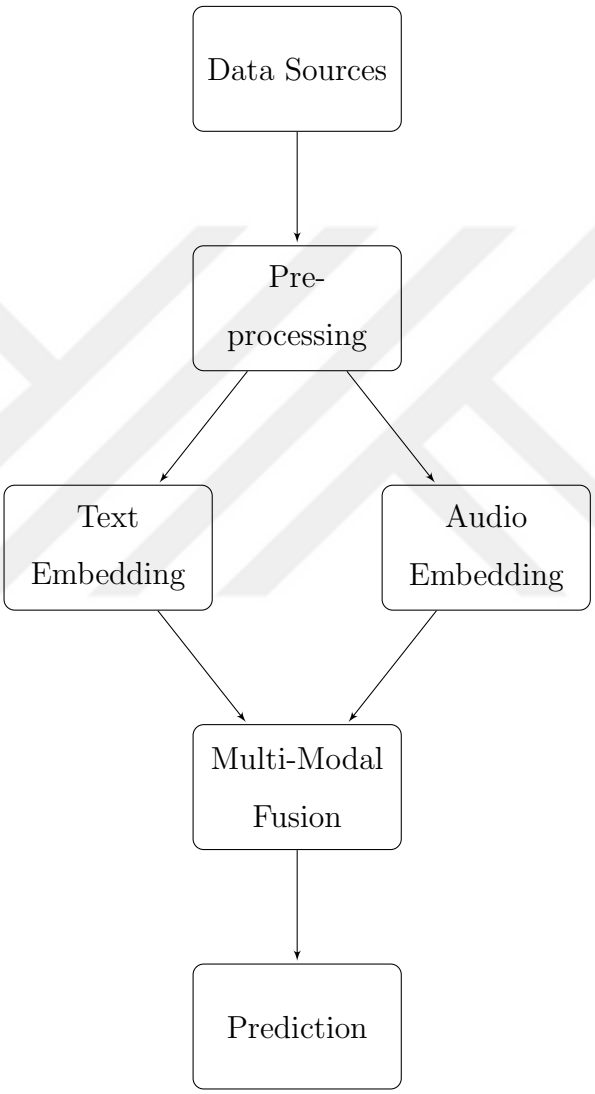



Figure 5.1: Overall Proposed Method for Backchannel Generation in Human-Robot Interaction

normalization is executed over discrete windows of voiced articulation that demonstrate pitch periodicity. The normalized intensity, pitch, and confidence-to-pitch parameters, along with their first temporal derivatives, constitute a 6-dimensional prosody feature vector. Finally, an extended 32-dimensional audio feature vector is assembled by concatenating the spectral and prosody features.

### *5.4.2 Textual Features*

Textual features are extracted from the transcriptions of speech present in the respective databases, as elaborated in Sections 5.3.2 and 5.3.3. These transcriptions serve as the basis for generating word-level embeddings, which are subsequently integrated into either feature-level fusion algorithms or unimodal prediction models. The methodology for windowing these word sequences during both the training and testing phases is delineated in following sections.

### *5.4.3 Cross-Attentional Multi-modal Transformers*

The core of the proposed method lies in the integration of Transformer models with cross-attention mechanisms. Transformers (Vaswani et al., 2017) provide parallelized attention across all input positions and have shown remarkable success in various NLP tasks. The cross-attention mechanism enables the model to attend to different parts of the input, allowing for a harmonized analysis.

Our proposed architecture focuses solely on the decoder part of the standard Transformer model, leveraging it to fuse multi-modal data—specifically textual and audio features. The overview of the model is depicted in Figure 5.2. The textual data is converted into a sequence of word embeddings, while the audio data is transformed into a sequence of feature vectors. Before entering the decoder, both the textual and audio sequences pass through self-attention layers to prepare the 'memory' that the decoder attends to. The self-attention mechanism is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5.1}$$

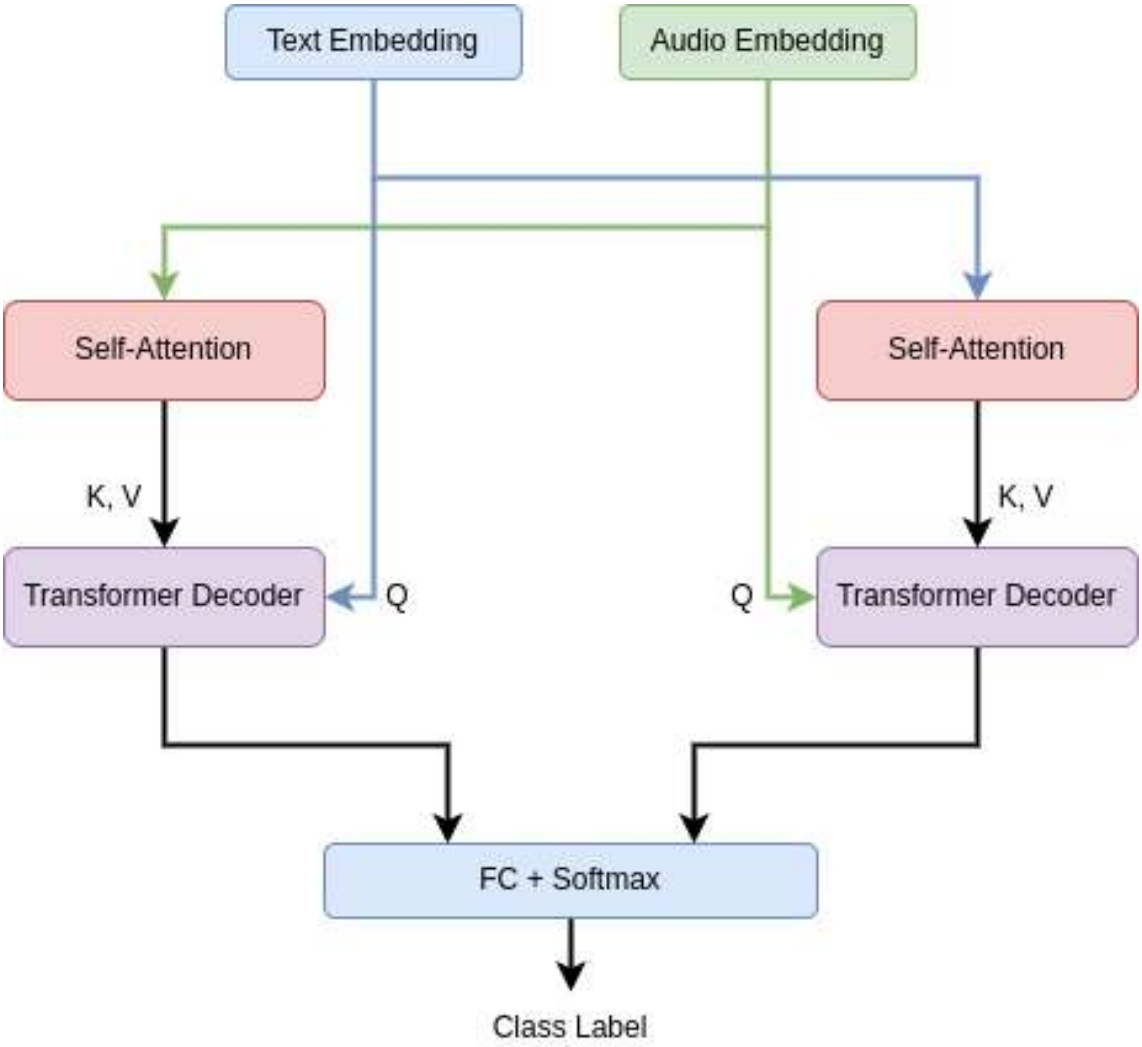



Figure 5.2: Overview of the Proposed Decoder-only Cross-Attention Model for Multi-modal Classification.

The core of our approach lies in the cross-attention layer within the decoder. We allow the textual modality to serve as the Query (Q) and the audio modality to serve as the Key (K) and Value (V). This cross-attention mechanism is described by:

$$\text{CrossAttention}(Q_{\text{T}}, K_{\text{A}}, V_{\text{A}}) = \text{softmax}\left(\frac{Q_{\text{T}} K_{\text{A}}^{T}}{\sqrt{d_k}}\right) V_{\text{A}} \tag{5.2}$$

where subscript T denotes text modality and A denotes audio modality.

The output of the cross-attention layer is further processed through additional self-attention and feed-forward layers. These layers aim to refine the cross-modal context and prepare it for final classification. Finally, the output of the last decoder layer serves as the input to a fully connected layer followed by a softmax activation function for classification:

$$\text{Class Labels} = \text{softmax}(W_f D_{Out}) \tag{5.3}$$

This model structure leverages the strengths of Transformer decoders to effectively fuse textual and audio modalities using cross-attention mechanisms, thus providing a robust solution for multi-modal classification tasks.

## *5.5 Experimental Results*

This section delineates the experimental procedures and outcomes, encompassing the fine-tuning of embedding models, training of classification models, and evaluation of uni-modal, baseline fusion, and cross-attentional multi-modal performances. The latter is a manifestation of the proposed methodology.

### *5.5.1 Setup*

In the experiments, OpenSubtitles data is used for fine-tuning text embedding model which is transformer encoder based BERT model. Similarly, wav2vec model is used to obtain audio embedding model. For the overall model training and testing, manually annotated Mimicry dataset is used.

The model ingests a context window as input and produces a binary classification output indicating the presence or absence of laughter immediately following the context window. Specifically, the model predicts the likelihood of laughter at time $t + T_p$ based on the observations within the window $[t - T_c, t]$. Figure 5.3 illustrates this configuration. The context window duration $T_c$ is set to 4 seconds for audio and 20 seconds for text, while the prediction lead time $T_p$ is set to 0.5 seconds to mitigate data leakage between the observation window and the label.

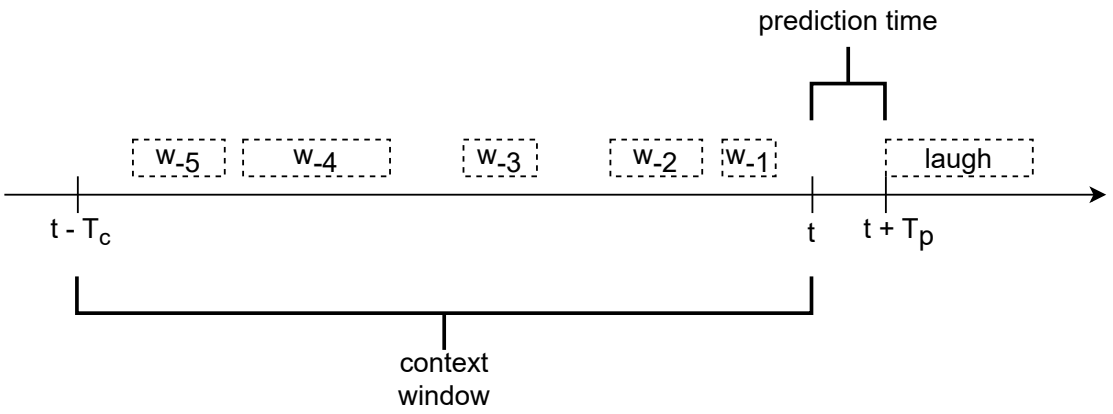



Figure 5.3: Prediction Configuration: $w$ denotes word sequences sub-scripted with time indices $t$, $T_c$ represents the context window duration, and $T_p$ indicates the prediction lead time for a potential laugh

Given the rarity of laughter events in natural conversations, the problem is characterized by a highly imbalanced class distribution. To augment the dataset with positive labels, multiple $T_p$ values, namely [0.5, 1.0, 1.5, 2.0], are employed. For the negative class, random sampling is conducted to balance the dataset.

### *5.5.2 Text Embedding Model*

The text embedding model undergoes supervised fine-tuning using the OpenSubtitles dataset, which contains labels related to laughter and smiles. Positive samples are constructed based on an approximate word count corresponding to a $T_c$ of 20 seconds, resulting in a 30-word observation window preceding the backchannel event. Negative samples are similarly generated but are randomly selected from regions that are safely distanced from positive samples.

The DistilBERT-base model, sourced from Huggingface, serves as the foundational transformer model. It is fine-tuned over three epochs using the OpenSubtitles dataset. Subsequently, the classification layer is removed, rendering the model suitable for extracting embeddings from word sequences.

### *5.5.3 Audio Embedding Model*

The audio embedding model employs a waveform as input to generate a feature sequence. The Mimicry dataset is used for fine-tuning the wav2vec2-based Hidden-Unit BERT (HuBERT) model. The pretrained model is sourced from Huggingface under the name Facebook-HuBERT-base. The performance of the audio uni-modal model is also assessed and reported in subsequent sections. Ultimately, the classification head is disregarded, and the last hidden layer output of the HuBERT model is utilized for embedding extraction from given audio waveforms.

### *5.5.4 Results*

The experiments employ balanced class-sampled datasets to evaluate and contrast the performance under various configurations and hyperparameters. One-fifth of the dataset is reserved for final testing, and the performance metrics are reported using precision, recall, $F_1 - score$ and AUC-ROC.

Contrastingly, the proposed methodology and the best-performing baseline are assessed using natural conversational clips, which inherently contain a preponderance of negative samples juxtaposed with rare positive instances.

#### *Results on Audio*

The audio unimodal prediction performance is ascertained during the fine-tuning phase of the audio embedding model. Early stopping is employed, monitoring the validation set accuracy over a maximum of six epochs. The learning rate incorporates a 10% warm-up ratio, and the batch size is set to 16. The resulting AUC-ROC performance on the Mimicry test set is 0.74.

#### *Results on Text*

The text embedding model undergoes a two-phase fine-tuning process. Initially, the model is fine-tuned using the OpenSubtitles dataset, which is notably extensive. Subsequently, a secondary fine-tuning is conducted on the Mimicry dataset. Various

configurations involving epochs and learning rates are empirically evaluated. The optimal model configuration is determined to have a fixed learning rate and is fine-tuned over five epochs on the OpenSubtitles dataset. This is followed by additional fine-tuning on the Mimicry dataset, also with a fixed learning rate, employing an early stopping mechanism to prevent overfitting. The resultant model achieves an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.71 on the Mimicry test set.

*Fusion Results*

In the context of fusion settings, each experiment employs feature-level fusion, succeeded by the training of a classification head. Feature-level fusion is essentially the concatenation of outputs from distinct modality-specific embedding models.

Table 5.1: Fusion models results

| **Experiment** | | | **Performance** | | | |
|---|---|---|---|---|---|---|
| **Fusion Type** | **Text** | **Audio** | **Pre** | **Rec** | $F_1$ | **AUC** |
| Feat. Level Concat. | Short | Short | 0.74 | 0.72 | 0.73 | 0.75 |
| Feat. Level Concat. | Long | Short | 0.75 | 0.73 | 0.74 | 0.77 |
| Cross Attention | Long | Short | 0.73 | 0.76 | 0.74 | 0.77 |
| Dual-way Cross Att. | Long | Short | 0.76 | 0.77 | 0.77 | 0.79 |

Table 5.1 presents the outcomes of four distinct experimental setups involving fusion models. It is noteworthy that the performance metrics are reported for a test set that is balanced with respect to instances of laughter and non-laughter. Observations indicate that extending the observation window for text modality leads to an enhancement in model performance. Additionally, the incorporation of a Cross Attention mechanism yields a significant improvement in predictive accuracy.

The Dual-way Cross Attention fusion approach utilizes the Query (Q), Key (K), and Value (V) inputs from the cross-attention mechanism in a bidirectional manner across modalities, followed by feature-level concatenation. In this setup, the

subscripts T and A in Equation 5.2 are interchanged, yielding an alternative model configuration that is hypothesized to offer a unique perspective on the data.

*Results on Full Clips*

As a foundational baseline, Long Short-Term Memory (LSTM) networks are utilized. These networks are proficient in capturing temporal dependencies and have garnered widespread application in the domain of sequential data modeling. The LSTM networks are configured with standard hyperparameters, thereby serving as a performance benchmark.

In the present analysis, unedited video clips are employed to simulate real-time prediction tasks. While previous sections focused on isolated prediction samples constituting class-balanced test sets, the current section feeds full-length clips into both the baseline LSTM and the proposed cross-attention transformer models. This approach facilitates a comprehensive performance comparison under conditions of high class imbalance.

For the training set, 36 out of 54 clips are retained, comprising 445 positive samples and a 30-fold larger set of randomly selected negative samples. The augmented negative sample selection aids the model in adapting to the highly imbalanced nature of the test set. For testing, the remaining 18 clips, totaling 3.74 hours, are processed using a sliding window technique. Positive samples are aggregated into groups of 8 consecutive frames, each spanning a duration of 4 seconds. This results in 246 laugh regions, generating 1968 positive samples. This aggregation strategy mitigates the timing precision requirements for the predictor, thereby facilitating a more equitable comparison.

Figure 5.4 displays the confusion matrix for the baseline LSTM model. While the number of true positives surpasses that of false negatives, a significant quantity of false positives is observed, closely rivaling the count of true negatives. Elevated false positive rates are a common phenomenon in such systems, often indicative of the model erroneously anticipating backchannel cues. It is worth noting that while the data may not manifest these cues, the potential for their occurrence could still

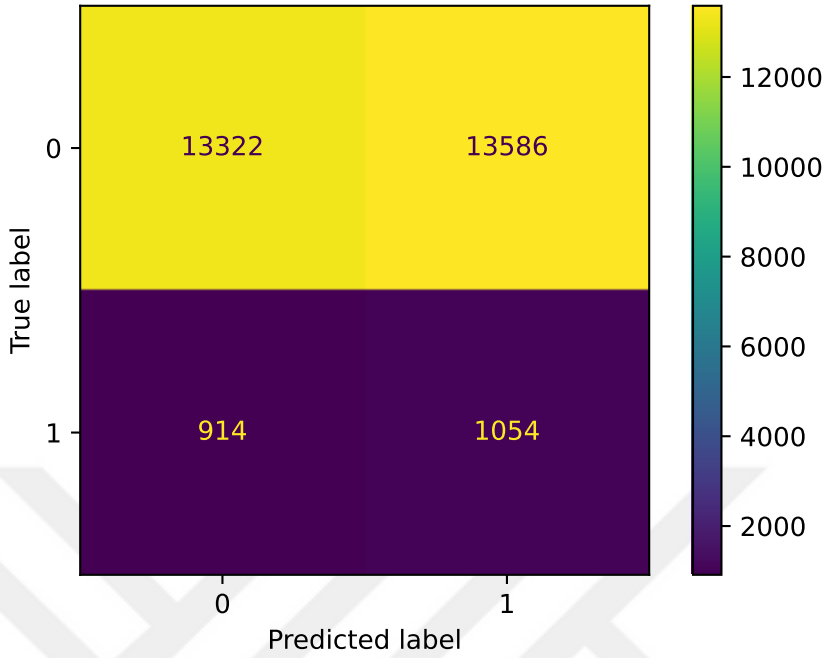



Figure 5.4: Baseline LSTM clip test results, confusion matrix. Pre:7.2%, Rec:53.6%, $F_1$: 12.7%

be present.

Figure 5.5 presents the confusion matrix for the performance of the proposed cross-attention transformer model. A noteworthy reduction in the false positive rate is observed, accompanied by an increase in the true positive rate. These outcomes are substantiated by the precision, recall, and $F_1$ score metrics, as delineated in the figure captions.

### *5.5.5 Discussion*

The experimental results offer several insights into the efficacy of various machine learning models for predicting laughter in dyadic interactions. The audio unimodal model yields an AUC-ROC of 0.74, while the text-based model slightly underperforms with an AUC-ROC of 0.71. This suggests that audio cues may be more indicative of imminent laughter than textual cues, although the difference is marginal.

The fusion models, particularly those employing cross-attention mechanisms,

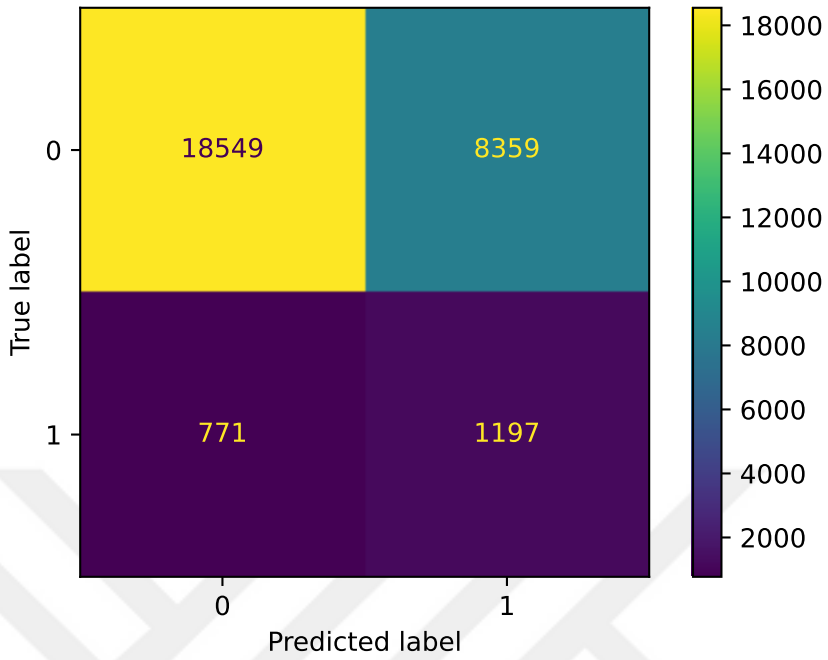



Figure 5.5: Cross Attention Transformer clip test results, confusion matrix. Pre:12.5%, Rec:60.8%, $F_1$: 20.7%

demonstrate superior performance, corroborating the utility of multi-modal approaches in capturing the nuances of human interaction. The dual-way cross-attention model exhibits the highest AUC-ROC of 0.79, indicating its robustness in handling the complexities of the task.

The baseline LSTM model, despite its capability to capture temporal dependencies, falls short in handling the highly imbalanced nature of the test set, as evidenced by its lower $F_1$ score. In contrast, the proposed cross-attention transformer model significantly reduces the false positive rate while enhancing the true positive rate, thereby validating its effectiveness in real-world scenarios.

Overall, the results affirm the potential of transformer-based models, especially those employing cross-attention mechanisms for multi-modal schemes, in advancing the field of Human-Robot Interaction by enabling more natural and engaging interactions.

## *5.6 Conclusion*

This research endeavored to develop a machine learning model for predicting laughter as backchannels in dyadic conversations, with a particular emphasis on the deployment of Transformer models in a multi-modal framework. The empirical findings corroborate the considerable efficacy of Transformer models equipped with cross-attention mechanisms, which exhibited superior performance compared to the LSTM baseline.

The study leveraged the Mimicry dyadic conversation database, enriched with manual annotations for smiles and laughter, thereby providing a robust dataset for experimental validation. By adopting distinct temporal windows for speech and acoustic cues, the model achieved a nuanced representation of the interactive dynamics inherent in dyadic conversations.

The key contributions of this research are as follows:

- **Innovative Application of Transformer Models:** The introduction of cross-attention Transformers to the field of Human-Robot Interaction (HRI) constitutes a novel contribution, demonstrating marked improvements in predictive accuracy.

- **Multi-modal Analytical Framework:** The integration of both speech and acoustic cues offers a holistic modeling approach, underscoring the necessity for a synchronized analysis of multiple communication modalities.

- **Comparative Evaluation:** A meticulous comparison with LSTM networks elucidates the intrinsic advantages of Transformer models in the specialized domain of non-verbal cue prediction.

The outcomes of this research bear significant ramifications for the conceptualization and engineering of socially interactive robotic systems. By enhancing the predictive accuracy of non-verbal cues, the potential for heightened engagement and naturalness in human-robot interactions is substantially increased, with applications spanning sectors such as education, healthcare, and entertainment.

While the research has made considerable strides, certain limitations warrant acknowledgment. The concentration on laughter as backchannels may not encapsulate the entire gamut of non-verbal communication in HRI. Future investigations could extend the model to encompass additional gestures and expressions and could also consider the impact of cultural variations.

Furthermore, the real-time deployment of the model and its personalization to individual users present intriguing prospects for future research, facilitating more adaptive and context-sensitive interactions.

In summary, this study lays a comprehensive groundwork for the predictive generation of laughter in social robots. By pioneering the application of Transformer models in this context, it makes a significant contribution to the burgeoning field of HRI. The methodologies and insights proffered herein serve as a catalyst for further scholarly inquiry, aimed at rendering human-robot interactions increasingly natural and engaging.

# Chapter 6

# CONCLUSION

This thesis has undertaken a comprehensive exploration into the realm of Human-Robot Interaction (HRI), with a particular focus on the prediction and utilization of non-verbal cues such as laughter, smiles, and head nods. The work has been grounded in the application of advanced machine learning techniques, including Support Vector Machines (SVMs), Time Delay Neural Networks (TDNNs), Long Short-Term Memory networks (LSTMs), and Transformer models. The research journey has been extensive, spanning from meticulous feature identification and algorithmic design to rigorous experimental validation.

One of the seminal contributions of this work lies in the domain of laughter detection. The thesis has addressed the challenges posed by continuous laughter detection in naturalistic dyadic conversations. Through a rigorous investigation of facial information, head movement, and audio features, the research has identified a set of discriminative features that significantly enhance the performance of classifiers. The introduction of bagging techniques to handle class imbalance has further strengthened the robustness of the laughter detection model, making it resilient to detrimental factors such as cross-talk and environmental noise.

Another pivotal contribution is the exploration of the role of laughter in terms of engagement in HRI. The thesis has designed and executed experiments that have provided both objective and subjective measures of engagement and user experience. The findings indicate that a laughter-responsive robot significantly enhances user engagement compared to a non-responsive counterpart, thereby underlining the importance of real-time laughter detection in HRI.

The thesis has also made significant strides in the prediction of head nods and turn-taking events, which are crucial for dialog management systems. By employing

SVMs and LSTMs, the research has developed an audio-visual prediction framework that is not only accurate but also amenable to real-time applications. The work has demonstrated the efficacy of these models on the IEMOCAP dataset, thereby contributing to the broader understanding of conversational dynamics in dyadic interactions.

Furthermore, the work has ventured into the prediction of upcoming candidate smiles and laughs as backchannels within dyadic conversations. By employing Transformer models with cross-attention mechanisms, the research has shown a significant improvement in prediction performance over traditional LSTM networks. This not only validates the efficacy of advanced deep learning techniques but also emphasizes the importance of a multi-modal approach in the modeling of backchannels in HRI.

While the thesis has made substantial contributions, it is imperative to acknowledge its limitations. The focus on a limited set of non-verbal cues and the reliance on specific machine learning models suggest avenues for future research. Extending the model to include other gestures and expressions, exploring different cultural contexts, and implementing real-time personalization are exciting directions for future work.

In conclusion, this thesis has laid a comprehensive foundation for the understanding and advancement of non-verbal cue prediction in HRI. It has significantly contributed to the evolving landscape of HRI research, offering new perspectives and methodologies that pave the way for further exploration and refinement. The insights garnered from this work hold substantial implications for the design and development of socially interactive robots, thereby opening up new horizons for the application of HRI in various domains such as education, healthcare, and entertainment.

# BIBLIOGRAPHY

[Akbani et al., 2004] Akbani, R., Kwek, S., and Japkowicz, N. (2004). *Applying Support Vector Machines to Imbalanced Datasets*, pages 39–50. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Al Moubayed et al., 2013] Al Moubayed, S., Beskow, J., and Skantze, G. (2013). The furhat social companion talking head. In *Interspeech 2013, 14th Annual Conference of the International Speech Communication Association, August 25-29, 2013, Lyon, France*, pages 747–749.

[Al Moubayed et al., 2012] Al Moubayed, S., Beskow, J., Skantze, G., and Granström, B. (2012). *Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction*, pages 114–130. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Andre et al., 2004] Andre, E., Rehm, M., Minker, W., and Bühler, D. (2004). Endowing spoken language dialogue systems with emotional intelligence. In *Tutorial and Research Workshop on Affective Dialogue Systems*, pages 178–187. Springer.

[Bickmore et al., 2011] Bickmore, T., Pfeifer, L., and Schulman, D. (2011). Relational agents improve engagement and learning in science museum visitors. In *International Workshop on Intelligent Virtual Agents*, pages 55–67. Springer.

[Bickmore and Picard, 2005] Bickmore, T. W. and Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293–327.

[Bilakhia et al., 2015] Bilakhia, S., Petridis, S., Nijholt, A., and Pantic, M. (2015).

The mahnob mimicry database: A database of naturalistic human interactions. *Pattern recognition letters*, 66:52–61.

[Bohus and Horvitz, 2014] Bohus, D. and Horvitz, E. (2014). Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th international conference on multimodal interaction*, pages 2–9. ACM.

[Bozkurt et al., 2015] Bozkurt, E., Erzin, E., and Yemez, Y. (2015). Affect-Expressive Hand Gestures Synthesis and Animation. In *IEEE International Conference on Multimedia and Expo (ICME)*, Torino, Italy.

[Bozkurt et al., 2016a] Bozkurt, E., Erzin, E., and Yemez, Y. (2016a). Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures. *Speech Communication*, 85:29–42.

[Bozkurt et al., 2016b] Bozkurt, E., Yemez, Y., and Erzin, E. (2016b). Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures. *Speech Communication*, 85:29–42.

[Breazeal, 2003] Breazeal, C. (2003). Emotion and sociable humanoid robots. *International journal of human-computer studies*, 59(1-2):119–155.

[Bryant and Aktipis, 2014] Bryant, G. A. and Aktipis, C. A. (2014). The animal nature of spontaneous human laughter. *Evolution and Human Behavior*, 35(4):327 – 335.

[Busso et al., 2008] Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

[Busso and Narayanan, 2008] Busso, C. and Narayanan, S. (2008). Recording audio-visual emotional databases from actors: A closer look. In *Second International*

*Workshop on Emotion: Corpora for Research on Emotion and Affect, International Conference on Language Resources and Evaluation*, pages 17–22, Marrakech, Morocco.

[Calisgan et al., 2012] Calisgan, E., Haddadi, A., Van der Loos, H. M., Alcazar, J. A., and Croft, E. A. (2012). Identifying nonverbal cues for automated human-robot turn-taking. In *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pages 418–423. IEEE.

[Carletta et al., 2006] Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., McCowan, I., Post, W., Reidsma, D., and Wellner, P. (2006). The ami meeting corpus: A pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction*, MLMI'05, pages 28–39, Berlin, Heidelberg. Springer-Verlag.

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27.

[Chang et al., 2022] Chang, S.-y., Li, B., Sainath, T. N., Zhang, C., Strohman, T., Liang, Q., and He, Y. (2022). Turn-taking prediction for natural conversational speech. *arXiv preprint arXiv:2208.13321*.

[Chartrand and Bargh, 1999] Chartrand, T. L. and Bargh, J. A. (1999). The chameleon effect: the perception–behavior link and social interaction. *Journal of personality and social psychology*, 76(6):893.

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

[Chen et al., 2015] Chen, Y., Yu, Y., and Odobez, J.-M. (2015). Head nod detection

from a full 3d model. In *Proceedings of the ICCV 2015*, number EPFL-CONF-213704.

[Choi et al., 2012] Choi, A., Melo, C. D., Woo, W., and Gratch, J. (2012). Affective engagement to emotional facial expressions of embodied social agents in a decision-making game. *Computer Animation and Virtual Worlds*, 23(3-4):331–342.

[Clavel et al., 2016] Clavel, C., Cafaro, A., Campano, S., and Pelachaud, C. (2016). *Fostering User Engagement in Face-to-Face Human-Agent Interactions: A Survey*, pages 93–120. Springer International Publishing, Cham.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

[Cosentino et al., 2016] Cosentino, S., Sessa, S., and Takanishi, A. (2016). Quantitative laughter detection, measurement, and classification - a critical survey. *IEEE Reviews in Biomedical Engineering*, 9:148–162.

[Cosker and Edge, 2009] Cosker, D. and Edge, J. (2009). Laughing, crying, sneezing and yawning: Automatic voice driven animation of non-speech articulations. *Proceedings of Computer Animation and Social Agents, CASA*.

[Dautenhahn, 2007] Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human–robot interaction. *Philosophical transactions of the royal society B: Biological sciences*, 362(1480):679–704.

[de Cheveigne¥ and Kawahara, 2002] de Cheveigne¥, A. and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917.

[Devillers et al., 2015] Devillers, L., Rosset, S., Duplessis, G. D., Sehili, M. A., Bechade, L., Delaborde, A., Gossart, C., Letard, V., Yang, F., Yemez, Y., Turker, B. B., Sezgin, M., Haddad, K. E., Dupont, S., Luzzati, D., Esteve, Y., Gilmartin,

E., and Campbell, N. (2015). Multimodal data collection of human-robot humorous interactions in the joker project. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 348–354.

[Douglas-Cowie. et al., 2003] Douglas-Cowie., E., Campbell, N., Cowie, R., and Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication*, 4(1-2):33–60.

[Dunbar, 2012] Dunbar, R. (2012). Bridging the bonding gap: the transition from primates to humans. *Philosophical Transactions of the Royal Society B Biological Sciences*, 367:1837–1846.

[Dybala et al., 2010] Dybala, M. P. P., Rzepka, R., and Araki, K. (2010). Forgetful and emotional: Recent progress in development of dynamic memory management system for conversational agents. In *Linguistic And Cognitive Approaches To Dialog Agents Symposium, Rafal Rzepka (Ed.), at the AISB 2010 convention.*

[Dybala et al., 2009] Dybala, P., Ptaszynski, M., Rzepka, R., and Araki, K. (2009). Activating humans with humor–a dialogue system that users want to interact with. *IEICE transactions on information and systems*, 92(12):2394–2401.

[D'Mello and Graesser, 2013] D'Mello, S. and Graesser, A. (2013). Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Trans Interact Intell Syst*, 4(2):1–39.

[Ekman et al., 1987] Ekman, P., Friesen, W. V., O'sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., et al. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53(4):712.

[El Haddad et al., 2016] El Haddad, K., Çakmak, H., Gilmartin, E., Dupont, S., and Dutoit, T. (2016). Towards a listening agent: a system generating audiovisual

laughs and smiles to show interest. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 248–255. ACM.

[Elfenbein, 2007] Elfenbein, H. A. (2007). 7 emotion in organizations: a review and theoretical integration. *The academy of management annals*, 1(1):315–386.

[Erzin et al., 2005] Erzin, E., Yemez, Y., and Tekalp, A. (2005). Multimodal speaker identification using an adaptive classifier cascade based on modality reliability. *IEEE Transactions on Multimedia*, 7(5):840–852.

[Escalera et al., 2009] Escalera, S., Puertas, E., Radeva, P., and Pujol, O. (2009). Multi-modal laughter recognition in video conversations. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 110–115.

[Fawcett, 2006] Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recogn. Lett.*, 27(8):861–874.

[Finger, 2013] Finger, S. (2013). Curious behavior: Yawning, laughing, hiccupping, and beyond by robert provine. *Journal of the History of the Neurosciences*, 22(4):429–430.

[Galar et al., 2012] Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. (2012). A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484.

[Glas and Pelachaud, 2015] Glas, N. and Pelachaud, C. (2015). Definitions of engagement in human-agent interaction. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*, pages 944–949.

[Gratch et al., 2007] Gratch, J., Wang, N., Gerten, J., Fast, E., and Duffy, R. (2007). *Creating Rapport with Virtual Agents*, pages 125–138. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Griffin et al., 2013] Griffin, H. J., Aung, M. S., Romera-Paredes, B., McLoughlin, C., McKeown, G., Curran, W., and Bianchi-Berthouze, N. (2013). Laughter type recognition from whole body motion. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 349–355. IEEE.

[Griffin et al., 2015] Griffin, H. J., Aung, M. S. H., Romera-Paredes, B., McLoughlin, C., McKeown, G., Curran, W., and Bianchi-Berthouze, N. (2015). Perception and automatic recognition of laughter from whole-body motion: Continuous and categorical perspectives. *IEEE Transactions on Affective Computing*, 6(2):165–178.

[Ho et al., 2015] Ho, S., Foulsham, T., and Kingstone, A. (2015). Speaking and listening with the eyes: gaze signaling during dyadic interactions. *PloS one*, 10(8):e0136905.

[Hochreiter and Schmidhuber, 1997a] Hochreiter, S. and Schmidhuber, J. (1997a). Long short-term memory. *Neural computation*, 9(8):1735–1780.

[Hochreiter and Schmidhuber, 1997b] Hochreiter, S. and Schmidhuber, J. (1997b). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

[Hofmann et al., 2017] Hofmann, J., Platt, T., and Ruch, W. (2017). Laughter and smiling in 16 positive emotions. *Ieee Transactions on Affective Computing*, 8:495–507.

[Inden et al., 2013] Inden, B., Malisz, Z., Wagner, P., and Wachsmuth, I. (2013). Timing and entrainment of multimodal backchanneling behavior for an embodied conversational agent. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 181–188, New York, NY, USA. ACM.

[Inoue et al., 2022] Inoue, K., Lala, D., and Kawahara, T. (2022). Can a robot laugh

with you?: shared laughter generation for empathetic spoken dialogue. *Frontiers in Robotics and Ai*, 9.

[Ishi et al., 2019] Ishi, C., Minato, T., and Ishiguro, H. (2019). Analysis and generation of laughter motions, and evaluation in an android robot. *Apsipa Transactions on Signal and Information Processing*, 8.

[Ito et al., 2005] Ito, A., Wang, X., Suzuki, M., and Makino, S. (2005). Smile and laughter recognition using speech processing and face recognition from conversation video. In *2005 International Conference on Cyberworlds (CW'05)*, pages 8 pp.–444.

[Jokinen et al., 2013] Jokinen, K., Furukawa, H., Nishida, M., and Yamamoto, S. (2013). Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(2):1–30.

[Kapoor and Picard, 2001] Kapoor, A. and Picard, R. W. (2001). A real-time head nod and shake detector. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–5. ACM.

[Kawahara et al., 2012] Kawahara, T., Iwatate, T., and Takanashi, K. (2012). Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations. In *Thirteenth Annual Conference of the International Speech Communication Association*.

[Kennedy and Ellis, 2004] Kennedy, L. S. and Ellis, D. P. (2004). Laughter detection in meetings.

[Kittler et al., 1998] Kittler, J., Hatef, M., Duin, R., and Matas, J. (1998). On combining classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(3):226–239.

[Knox and Mirghafori, 2007] Knox, M. T. and Mirghafori, N. (2007). Automatic laughter detection using neural networks. In *Interspeech*, pages 2973–2976.

[Krikke and Truong, 2013] Krikke, T. F. and Truong, K. P. (2013). Detection of nonverbal vocalizations using gaussian mixture models: looking for fillers and laughter in conversational speech. In *Interspeech*, pages 163–167.

[Krumhuber and Scherer, 2011] Krumhuber, E. and Scherer, K. R. (2011). Affect bursts: Dynamic patterns of facial expression. *Emotion*, 11:825–841.

[Laskowski and Schultz, 2008] Laskowski, K. and Schultz, T. (2008). Detection of laughter-in-interaction in multichannel close-talk microphone recordings of meetings. In *Machine Learning for Multimodal Interaction*, pages 149–160. Springer.

[Lee and Marsella, 2010] Lee, J. and Marsella, S. C. (2010). Predicting speaker head nods and the effects of affective information. *IEEE Transactions on Multimedia*, 12(6):552–562.

[Li et al., 2020] Li, J., Ji, D., Li, F., Zhang, M., and Liu, Y. (2020). Hitrans: A transformer-based context-and speaker-sensitive model for emotion detection in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4190–4200.

[Lison and Tiedemann, 2016] Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles.

[Ma et al., 2019] Ma, J., Tang, H., Zheng, W.-L., and Lu, B.-L. (2019). Emotion recognition using multimodal residual lstm network. In *Proceedings of the 27th ACM international conference on multimedia*, pages 176–183.

[McKeown et al., 2012] McKeown, G., Cowie, R., Curran, W., Ruch, W., and Douglas-Cowie, E. (2012). Ilhaire laughter database. In *ES$^3$ 2012 4th International Workshop on Corpora for Research on emotion, sentiment, & social signals at the eighth international conference on Language Resources and Evaluation (LREC)*.

[Meena et al., 2014] Meena, R., Skantze, G., and Gustafson, J. (2014). Data-driven models for timing feedback responses in a map task dialogue system. *Computer Speech & Language*, 28(4):903–922.

[Mehrabian and Epstein, 1972] Mehrabian, A. and Epstein, N. (1972). A measure of emotional empathy. *Journal of personality*.

[Metallinou et al., 2013] Metallinou, A., Katsamanis, A., and Narayanan, S. (2013). Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing*, 31(2):137–152.

[Morency et al., 2008] Morency, L.-P., De Kok, I., and Gratch, J. (2008). Predicting listener backchannels: A probabilistic multimodal approach. In *International Workshop on Intelligent Virtual Agents*, pages 176–190. Springer.

[Morency et al., 2010] Morency, L.-P., de Kok, I., and Gratch, J. (2010). A probabilistic multimodal approach for predicting listener backchannels. *Autonomous Agents and Multi-Agent Systems*, 20(1):70–84.

[Napier and Leeson, 2016] Napier, J. and Leeson, L. (2016). Sign language in action. In *Sign Language in Action*, pages 50–84. Springer.

[Niewiadomski et al., 2013a] Niewiadomski, R., Hofmann, J., Urbain, J., Platt, T., Wagner, J., Bilal, P., Ito, T., Jonker, C., Gini, M., and Shehory, O. (2013a). Laugh-aware virtual agent and its impact on user amusement.

[Niewiadomski et al., 2013b] Niewiadomski, R., Hofmann, J., Urbain, J., Platt, T., Wagner, J., Piot, B., Cakmak, H., Pammi, S., Baur, T., Dupont, S., Geist, M., Lingenfelser, F., McKeown, G., Pietquin, O., and Ruch, W. (2013b). Laugh-aware virtual agent and its impact on user amusement. In *Proceedings of the 2013*

*International Conference on Autonomous Agents and Multi-agent Systems*, AAMAS '13, pages 619–626, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

[Niewiadomski et al., 2013c] Niewiadomski, R., Mancini, M., Baur, T., Varni, G., Griffin, H., and Aung, M. S. H. (2013c). *MMLI: Multimodal Multiperson Corpus of Laughter in Interaction*, pages 184–195. Springer International Publishing, Cham.

[Niewiadomski et al., 2016] Niewiadomski, R., Mancini, M., Varni, G., Volpe, G., and Camurri, A. (2016). Automated laughter detection from full-body movements. *IEEE Transactions on Human-Machine Systems*, 46(1):113–123.

[Novick et al., 1996] Novick, D. G., Hansen, B., and Ward, K. (1996). Coordinating turn-taking with gaze. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1888–1891 vol.3.

[Paiva et al., 2017] Paiva, A., Leite, I., Boukricha, H., and Wachsmuth, I. (2017). Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(3):1–40.

[Pecune et al., 2015] Pecune, F., Mancini, M., Biancardi, B., Varni, G., Ding, Y., Pelachaud, C., Volpe, G., and Camurri, A. (2015). Laughing with a virtual agent. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, pages 1817–1818, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

[Peng et al., 2005] Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238.

[Peters et al., 2009] Peters, C., Castellano, G., and de Freitas, S. (2009). An exploration of user engagement in hci. In *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*, AFFINE '09, pages 9:1–9:3, New York, NY, USA. ACM.

[Petridis et al., 2013a] Petridis, S., Leveque, M., and Pantic, M. (2013a). Audiovisual detection of laughter in human-machine interaction. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 129–134. IEEE.

[Petridis et al., 2013b] Petridis, S., Martinez, B., and Pantic, M. (2013b). The mahnob laughter database. *Image Vision Comput.*, 31(2):186–202.

[Petridis and Pantic, 2008] Petridis, S. and Pantic, M. (2008). Fusion of audio and visual cues for laughter detection. In *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*, CIVR '08, pages 329–338, New York, NY, USA. ACM.

[Petridis and Pantic, 2011] Petridis, S. and Pantic, M. (2011). Audiovisual discrimination between speech and laughter: Why and when visual information might help. *Multimedia, IEEE Transactions on*, 13(2):216–234.

[Petridis et al., 2011] Petridis, S., Pantic, M., and Cohn, J. F. (2011). Prediction-based classification for audiovisual discrimination between laughter and speech. In *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, pages 619–626.

[Poppe et al., 2014] Poppe, R., ter Maat, M., and Heylen, D. (2014). Switching wizard of oz for the online evaluation of backchannel behavior. *Journal on Multimodal User Interfaces*, 8(1):109–117.

[Poppe et al., 2013] Poppe, R., Truong, K. P., and Heylen, D. (2013). Perceptual

evaluation of backchannel strategies for artificial listeners. *Autonomous Agents and Multi-Agent Systems*, 27(2):235–253.

[Prylipko et al., 2012] Prylipko, D., Schuller, B., and Wendemuth, A. (2012). Fine-tuning hmms for nonverbal vocalizations in spontaneous speech: A multicorpus perspective. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4625–4628.

[Reuderink et al., 2008] Reuderink, B., Poel, M., Truong, K., Poppe, R., and Pantic, M. (2008). Decision-level fusion for audio-visual laughter detection. In *Machine Learning for Multimodal Interaction*, pages 137–148. Springer.

[Rich et al., 2010] Rich, C., Ponsler, B., Holroyd, A., and Sidner, C. L. (2010). Recognizing engagement in human-robot interaction. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 375–382.

[Ruch, 2008] Ruch, W. (2008). Psychology of humor. *The primer of humor research*, 8:17–100.

[Ruch and Ekman, 2001] Ruch, W. and Ekman, P. (2001). The Expressive Pattern of Laughter. *Emotion qualia, and consciousness*, pages 426–443.

[Scherer, 1994] Scherer, K. R. (1994). *Affect Bursts*. Emotions: Essays on Emotion Theory. Taylor & Francis Group.

[Scherer et al., 2012] Scherer, S., Glodek, M., Schwenker, F., Campbell, N., and Palm, G. (2012). Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data. *ACM Trans. Interact. Intell. Syst.*, 2(1):4:1–4:31.

[Schroder et al., 2012] Schroder, M., Bevacqua, E., Cowie, R., Eyben, F., Gunes, H., Heylen, D., Ter Maat, M., McKeown, G., Pammi, S., Pantic, M., et al. (2012).

Building autonomous sensitive artificial listeners. *IEEE Transactions on Affective Computing*, 3(2):165–183.

[Schuller et al., 2008] Schuller, B., Eyben, F., and Rigoll, G. (2008). Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech. In *Perception in multimodal dialogue systems*, pages 99–110. Springer.

[Schuller et al., 2013] Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., et al. (2013). The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

[Sidner et al., 2005a] Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., and Rich, C. (2005a). Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164.

[Sidner et al., 2005b] Sidner, C. L., Lee, C., Kidd, C. D., Lesh, N., and Rich, C. (2005b). Explorations in engagement for humans and robots. *Artif. Intell.*, 166(1-2):140–164.

[Skantze, 2017] Skantze, G. (2017). Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 220–230.

[Skantze, 2021] Skantze, G. (2021). Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*, 67:101178.

[Skantze and Al Moubayed, 2012] Skantze, G. and Al Moubayed, S. (2012). Iristk: A statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pages 69–76, New York, NY, USA. ACM.

[Skantze et al., 2014] Skantze, G., Hjalmarsson, A., and Oertel, C. (2014). Turn-taking, feedback and joint attention in situated human–robot interaction. *Speech Communication*, 65:50–66.

[Suarez et al., 2012] Suarez, M. T., Cu, J., and Maria, M. S. (2012). Building a multimodal laughter database for emotion recognition. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

[Szameitat et al., 2022] Szameitat, D., Szameitat, A., and Wildgruber, D. (2022). Vocal expression of affective states in spontaneous laughter reveals the bright and the dark side of laughter. *Scientific Reports*, 12.

[Truong and Leeuwen, 2005] Truong, K. P. and Leeuwen, D. A. v. (2005). Automatic detection of laughter. In *Ninth European Conference on Speech Communication and Technology*.

[Truong and Leeuwen, 2007] Truong, K. P. and Leeuwen, D. A. V. (2007). Automatic discrimination between laughter and speech. *Speech Communication*, 49(2):144–158.

[Truong et al., 2010] Truong, K. P., Poppe, R., and Heylen, D. (2010). A rule-based backchannel prediction model using pitch and pause information.

[Truong et al., 2021] Truong, T.-D., Duong, C. N., Pham, H. A., Raj, B., Le, N., Luu, K., et al. (2021). The right to talk: An audio-visual transformer approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1105–1114.

[Türker et al., 2017a] Türker, B. B., Buçinca, Z., Erzin, E., Yemez, Y., and Sezgin, M. T. (2017a). Analysis of engagement and user experience with a laughter responsive social robot. In *18th Annual Conference of the International Speech Communication Association*, Stockholm, Sweden.

[Türker et al., 2017b] Türker, B. B., Buçinca, Z., Erzin, E., Yemez, Y., and Sezgin, M. T. (2017b). Real-time audiovisual laughter detection. In *2017 25th Signal Processing and Communications Applications Conference (SIU)*.

[Türker et al., 2014] Türker, B. B., Marzban, S., Erzin, E., Yemez, Y., and Sezgin, T. M. (2014). Affect burst recognition using multi-modal cues. In *Signal Processing and Communications Applications Conference (SIU), 2014 22nd*, pages 1608–1611. IEEE.

[Turker et al., 2015] Turker, B. B., Marzban, S., Sezgin, M. T., Yemez, Y., and Erzin, E. (2015). Affect burst detection using multi-modal cues. In *2015 23nd Signal Processing and Communications Applications Conference (SIU)*, pages 1006–1009.

[Turker et al., 2017] Turker, B. B., Yemez, Y., Sezgin, T. M., and Erzin, E. (2017). Audio-facial laughter detection in naturalistic dyadic conversations. *IEEE Transactions on Affective Computing*, 8(4):534–545.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[Waibel et al., 1989] Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. J. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(3):328–339.

[Wittenburg et al., 2006] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: A professional framework for multimodality research. In *5th international conference on language resources and evaluation (LREC 2006)*, pages 1556–1559.

[Xie et al., 2019] Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C., and Schuller,

B. (2019). Speech emotion classification using attention-based lstm. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11):1675–1685.

[Yngve, 1970] Yngve, V. H. (1970). On getting a word in edgewise. In *Chicago Linguistics Society, 6th Meeting*, pages 567–578.

[Zeng et al., 2009] Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2009). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58.

[Zhao and Allison, 2017] Zhao, J. and Allison, R. S. (2017). Real-time head gesture recognition on head-mounted displays using cascaded hidden markov models. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2361–2366.

[Zhong et al., 2019] Zhong, P., Wang, D., and Miao, C. (2019). Knowledge-enriched transformer for emotion detection in textual conversations. *arXiv preprint arXiv:1909.10681*.

[Zou et al., 2022] Zou, S., Huang, X., Shen, X., and Liu, H. (2022). Improving multimodal fusion with main modal transformer for emotion recognition in conversation. *Knowledge-Based Systems*, 258:109978.

[Zue and Glass, 2000] Zue, V. and Glass, J. (2000). Conversational interfaces: Advances and challenges. *Proc. IEEE*, 42:1166–1180.