Audio-Facial Laughter Detection in Naturalistic Dyadic Conversations

Bekir Berker Turker, Yucel Yemez, T. Metin Sezgin, and Engin Erzin, *Senior Member, IEEE*

Abstract—We address the problem of continuous laughter detection over audio-facial input streams obtained from naturalistic dyadic conversations. We first present meticulous annotation of laughters, cross-talks and environmental noise in an audio-facial database with explicit 3D facial mocap data. Using this annotated database, we rigorously investigate the utility of facial information, head movement and audio features for laughter detection. We identify a set of discriminative features using mutual information-based criteria, and show how they can be used with classifiers based on support vector machines (SVMs) and time delay neural networks (TDNNs). Informed by the analysis of the individual modalities, we propose a multimodal fusion setup for laughter detection using different classifier-feature combinations. We also effectively incorporate bagging into our classification pipeline to address the class imbalance problem caused by the scarcity of positive laughter instances. Our results indicate that a combination of TDNNs and SVMs lead to superior detection performance, and bagging effectively addresses data imbalance. Our experiments show that our multimodal approach supported by bagging compares favorably to the state of the art in presence of detrimental factors such as cross-talk, environmental noise, and data imbalance.

Index Terms—Laughter detection, naturalistic dyadic conversations, facial mocap, data imbalance

1 Introduction

Laughter serves as an expressive social signal in human communication, and conveys distinctive information on 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 the design of more natural and human-centered interfaces with better user engagement [1].

Laughter is primarily a nonverbal vocalization accompanied with body and facial movements [2]. The majority of the existing automatic laughter detection methods in the literature have focused on audio-only information [3], [4]. This is mainly because audio is relatively easier to capture and analyze compared to 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 [5], [6], there are so far few works that exploit facial motion [3]. The main bottleneck here, especially for incorporation of 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 laughter. 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 paper is to investigate facial and head movements for their use in laughter detection over a multimodal database that comprises facial 3D motion capture data along with audio.

We address several challenges involved in 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 along with 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 presence of cross-talks and environmental noise.

2 Related Work

The work presented here falls under the general field of affective computing. However laughter detection is very

• *The authors are with the College of Engineering, Koc University, Istanbul 34450, Turkey. E-mail: {bturker13, yyemez, mtsezgin, eerzin}@ku.edu.tr.*

Manuscript received 29 July 2016; revised 17 Aug. 2017; accepted 10 Sept. 2017. Date of publication 18 Sept. 2017; date of current version 5 Dec. 2017. (Corresponding author: Bekir Berker Turker.)
Recommended for acceptance by M. Mancini, R. Niewiadomski, S. Hashimoto, M. E. Foster, S. Scherer, and G. Volpe.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TAFFC.2017.2754256

different from affect recognition [7] 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 [3], [8], [9], [10], [11], [12]. Laughter falls under the category of 'affect bursts' which denote *specific identifiable events* [13]. 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 [8], and Laskowski et al. explored laughter detection in the context of meetings [9]. Other unimodal work focused on body movements [5], [6]. Griffin et al. [6] studied how laughter is perceived by humans through avatar-based perceptual studies, and also explored automatic recognition of laughter from body movements. In another work, Griffin et al. [14] 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. [5] 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 a human performance.

To distinguish pre-segmented non-verbal vocalizations using audio only features, solutions employing different learning methods have been proposed. Schuller et al. [15] used a variety of classifiers on dynamic and static representations to differentiate non-verbal vocalizations (laughter, breathing, hesitation, and consent). Among 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 into how audio and motion information contribute to laughter recognition as individual modalities.

There are also other lines of work that have indirectly studied laughter detection on the course of addressing other problems. For example, since laughter is treated as noise in speech recognition, its detection, segmentation, and suppression have received attention [16]. Here laughter detection is our primary concern, and we treat it with rigor.

Our work is more closely related to the multimodal laughter detection and recognition systems [3], [10], [11], [17], [18], [19]. Escalera et al. combined audio information with smile-laughter events detected at the frame level and identified regions of laughter [17]. Petridis et al. proposed methods for discrimination between pre-segmented laughter and speech using both audio and video streams [3], [18]. Extracted features are, facial expressions and head pose from video, and cepstral and prosodic features from audio. They have also showed 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. [11] 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 clear advantage through the use of bagging for dealing with imbalanced databases.

In other multimodal work, Cosker and Edge [20] present analysis of correlation between voice and facial marker points using HMMs in four non-speech articulations, namely laughing, crying, sneezing and yawning. Although their work is geared towards synthesis of facial movements using sound, it provides useful insights into laughter recognition as well. A recent study done by Krumhuber and Scherer [21] shows that the facial action units, coded using Facial Action Coding System (FACS), exhibit significant variations for different affect bursts and hence can serve as cues in detecting and recognizing laughters.

Scherer et al. [22] proposed a multimodal laughter detection system based on Support Vector Machines, Echo State Networks and Hidden Markov Models. Although they use body and head movement information, the information is extracted from video. Similarly, Reuderink et al. [10] perform audiovisual laughter recognition on a modified and re-annotated version of the AMI Meeting Corpus [23]. 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 [24]. Although this method is promising, it could not directly be used for continuous detection of laughters 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 [25], [26], [27], [28], 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 were able to train models with training data selected based on their clean, noisy and/or cross-talk labels.

2.1 Contributions

In view of the related work discussed previously, the contributions of this paper 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

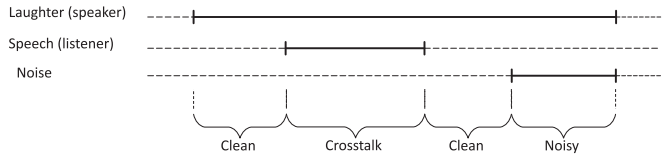

Fig. 1. Sample annotation fragment with speaker laughter, listener speech and noise.

information so as 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 the 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 laughters and demonstrate the advantage of incorporating facial and head features, especially to handle cross-talks and environmental noise.

3 Database

In our analysis and experimental evaluations, we have used the interactive emotional dyadic motion capture database (IEMOCAP), which is designed to study expressive human interactions [29]. The IEMOCAP is an audio-facial database, which provides motion capture data of face, head and partially hands as well as speech. The corpus has five sessions with ten professional actors taking part in dyadic interactions. Each session comprises spontaneous conversations under eight hypothetical scenarios as well as three scripted plays in order 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 laughters, and hence the laughter occurrences in the database are not pre-planned in any of the 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 [30]. The genuineness of acted emotions is an open issue in most of the existing "semi-natural" databases as also mentioned in this paper. Yet several works exist in the literature, which are addressing this issue and suggesting possible strategies to increase genuineness of acted emotions [30], [31], [32]. 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 which includes accurate facial measurements in the case of fully natural laughters.

In the IEMOCAP database, a VICON motion capture system is used to track 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 speaker and the other actor as listener. Speakers' data will be in the main focus of this study, as they include both audio and facial marker information. However, listeners have also impact on the audio channel by creating cross-talk effect on speakers' laughters. The authors of [29] note that the markers attached to speakers during motion data acquisition are very small so as not to interfere with natural speech. According to [29], the subjects also confirmed that they were comfortable with the markers, which did not prevent them from speaking naturally.

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

TABLE 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

3.1 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 footsteps of recording crew, 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 which is 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 Fig. 1.

All five sessions of the IEMOCAP database have been annotated as described above. Table 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 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 to 213.92 sec. The average duration of the laughter segments also decreases

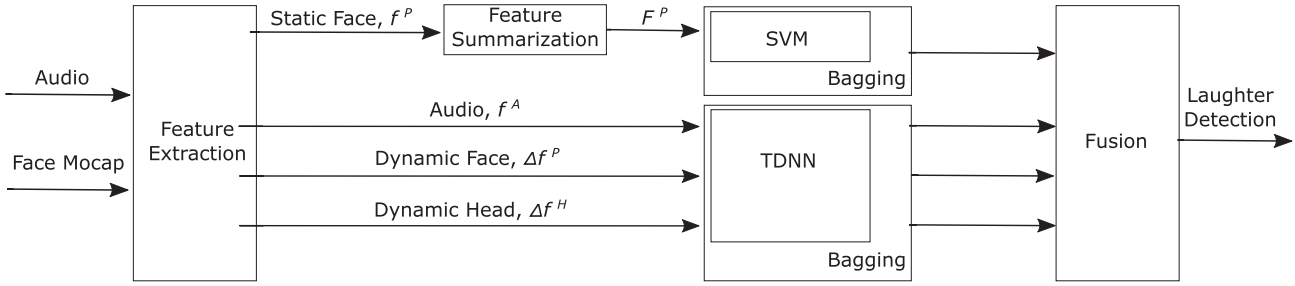

Fig. 2. Block diagram of the proposed laughter detection system. The diagram shows the best performing setup for classifier-feature combinations.

from 1.54 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 1, the total duration of clean laughter sequences is 213.92 sec, while the annotated IEMOCAP database is 8 hours. This causes an data imbalance between the two event classes of interest, laughter and non-laughter. To train our classifiers, we construct a balanced database (BLD) 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 so as to match the total duration of laughters.

4 Methodology

The main objective in 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 Fig. 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 and time-delay neural networks, receive a temporal window of short-term features or their summarizations in order to perform laughter versus non-laughter binary classification. The classification task is repeated for 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 so as 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.

4.1 Audio Features

Acoustic laughter signals can be characterized by their spectral properties as well as their prosodic structures. 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 laughters [3]. 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 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 [33], [34]. 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 [35]. Confidence-to-pitch delivers an auto-correlation score for the fundamental frequency of the signal [34].

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: $f^A = [f^M f^S]$.

4.2 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 [29]. We will refer to f^H as static head features, whereas dynamic head features will be represented by Δf^H, which are simply the first derivatives of static features. We note that dynamic head features are less dependent to 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 [5], [6] calculate head related features based on positioning of head with respect to full body such as the distance between shoulders and head as in [5]. 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 facial features and dynamic facial features. 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 tip of nose in the IEMOCAP database [29]. We denote the static facial feature vector by f^P and the dynamic facial feature vector by Δf^P which is the first derivative of the static version. We note that pose invariance is less of a problem in this case compared to head motion since facial points are compensated for rigid motion.

4.3 Feature Summarization

We use feature summarization for temporal characterization of laughter. For feature summarization, we compute a set of statistical quantities that describe the short-term distribution of each frame-level feature over a given window. 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. [36] for continuous tracking of emotional states from speech and body motion. We denote the window-level statistical features resulting from 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, ΔF^H, F^P and ΔF^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 use these window-level statistical features later to feed SVM classifiers for laughter detection.

4.4 Discriminative Analysis of Facial Laughter

In this section, we perform discriminative analysis on facial points to determine the relevance of each point in formation of laughter expression over the given database. We also explore possible correlations between facial points in order to eliminate redundancies in distinguishing laughter. Such correlations exist especially between symmetrical facial points (e.g., right cheek versus left cheek) and between points belonging 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 minimum Redundancy Maximum Relevance (mRMR)[37]. This method assigns an importance score to each feature, which measures its relevance to target classes under minimum redundancy condition. Relevance is defined based on mutual information between features and the corresponding class labels (laughter versus 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 as Maximum Relevance (maxRel)[37], 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 in order to observe importance of individual points for laughter, and mRMR to find optimal subsets of features to be fed into our classification engine.

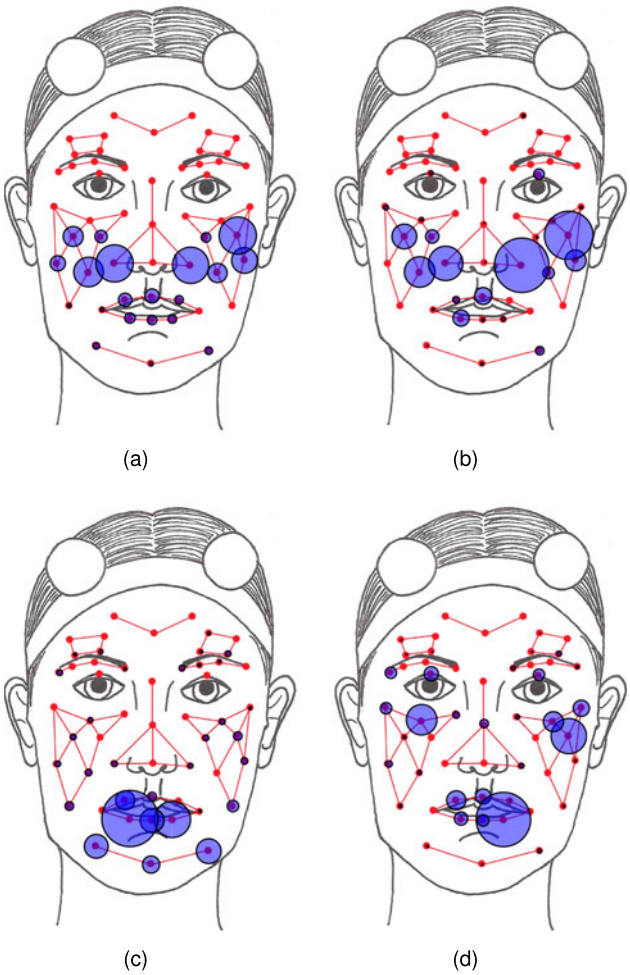

Fig. 3. Visualization of importance of facial points for laughter using (a) maxRel-static features, (b) mRMR-static features, (c) maxRel-dynamic features, and (d) mRMR - dynamic features. The size of a disc is proportional to importance of its center point.

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 3×11 statistical features in our case, the complete list has $53 \times 3 \times 11$ features, where 53 is the number of facial markers. To quantify the importance of each facial point based on this ordered list of features with individual scores, we employ a voting technique. 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. Figs. 3a and 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 which is focused

on certain regions of the face, especially on 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.

Figs. 3c and 3d display the results of maxRel and mRMR for dynamic facial features ΔF^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 relatively good amount of movement than any other regions of the face. In mRMR results however, we see that the points over chin region no longer appears 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 Fig. 3d, we also observe that points around the eyes start to get more importance, which is an indication of genuineness of the laughters in the database [2].

4.5 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 [38]. SVM projects data samples to different spaces through kernels that range from simple linear to radial basis function (RBF) [39]. 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 [40]. 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 has been successfully used by Petridis et al. [11]. 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 5.

4.5.1 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 3.1. This problem has been addressed and several solutions have been suggested in the literature [41]. For instance, SVM classifier performs better when class samples are balanced in the training [42]. Otherwise, it tends to favor the majority class. To deal with this problem, several methods have been proposed, 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 3.1. 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 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 [43], [44]. 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 5.

4.5.2 Fusion

The last block of Fig. 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 $f_1, f_2, \ldots, f_N$, are available, and for the nth 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

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

where ω_n denotes the weighting coefficient for classifier n, such that $\sum_n \omega_n = 1$. Note that when $\omega_n = \frac{1}{N}\ \forall n$, (1) is equivalent to the product rule [43], [44]. In this study, we employ the product rule and set all weights equal for decision fusion of the multimodal classifiers.

5 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 leave-one-session-out fashion, which results in a 5-fold train/test performance analysis. Since subjects are different across sessions, the reported results are also speaker independent. In each fold, training is carried out over a realization of the BLD, which is 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 4, laughter detection is performed as a classification task for every 250 msec over temporal windows of length 750 msec. Any temporal window, which contains a laughter segment longer than 375 msec, is taken as a laughter event.

For the SVM classifiers, the RBF kernel is used for all modalities with the hyper-parameters c and γ. 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 γ parameters so as 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, number of hidden nodes and time-delay. Since TDNN

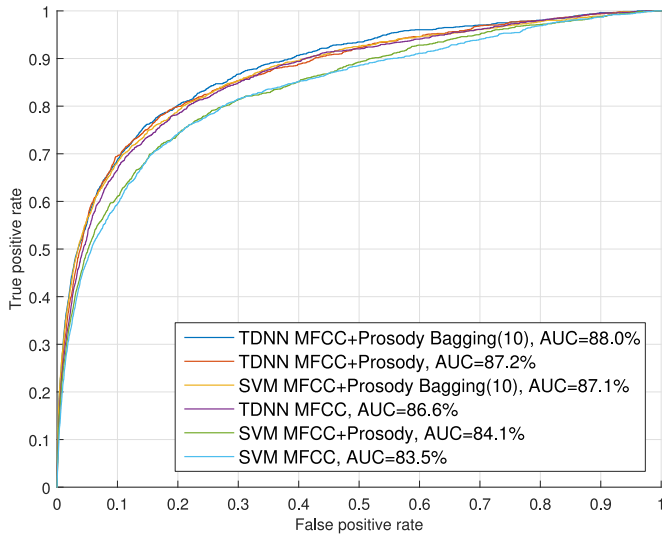

Fig. 4. Audio laughter detection ROC curves and AUC performances of the SVM and TDNN classifiers with different feature sets and with/without bagging.

classifiers exhibit increasing performance as number of hidden nodes and time-delay increase, we tend to set these parameters as low as possible to minimize the computational complexity (for especially training phase) while maintaining a high classification performance.

Finally, the resulting laughter detection performances are reported in terms of area under curve (AUC) percentage of the receiver operating characteristic (ROC) curves, which represents the overall performance over all operating points of the classifier [45]. Note that AUC is 100 percent for the ideal classifier and 50 percent for a random classifier.

5.1 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 4 as in [11]. We also apply 10-fold bagging to the best performing classifiers.

Fig. 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 percent AUC performance improvement, whereas the TDNN classifier improves 0.8 percent 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 percent AUC performance.

5.2 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 (τ_d), are set for the motion features by testing a fixed number of configurations with $\tau_d = 4, 8, 16$ and $n_h = 5, 10, 20$.

TABLE 2
The AUC Performances (%) of TDNN Classifiers with Dynamic Head Features Δf^H and with Varying Delay τ_d and Number of Hidden Nodes n_h

	Hidden Nodes (n_h)		
Delay (τ_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

In our preliminary studies, the static head features, f^H and F^H, have performed poorly for laughter detection, possibly due to the fact that these features have severe pose invariance problems. Hence in this paper, we consider only the dynamic features Δf^H and ΔF^H for head motion representation. The laughter detection performances of the TDNN classifiers for varying τ_d and n_h values are given in Table 2. We observe that TDNN achieves the best AUC performance as 87.2 percent with parameters $\tau_d = 16$ and $n_h = 20$. Yet, the other AUC performances with 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 percent AUC performance. On the other hand, the SVM classifier attains 81.5 percent 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.

For laughter detection from facial motion, we consider both static and dynamic features, f^P, F^P, Δf^P and ΔF^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 4.4, 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 versus dynamic and TDNN versus SVM) with varying number of features selected based on the mRMR discriminative analysis. In Fig. 5, we plot the AUC performances of these combinations with varying feature dimensions.

In Fig. 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 percent at feature dimension 23. Its performance saturates at this point and then starts to degrade slightly with 96.9, 94.8 and 95.2 percent 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.

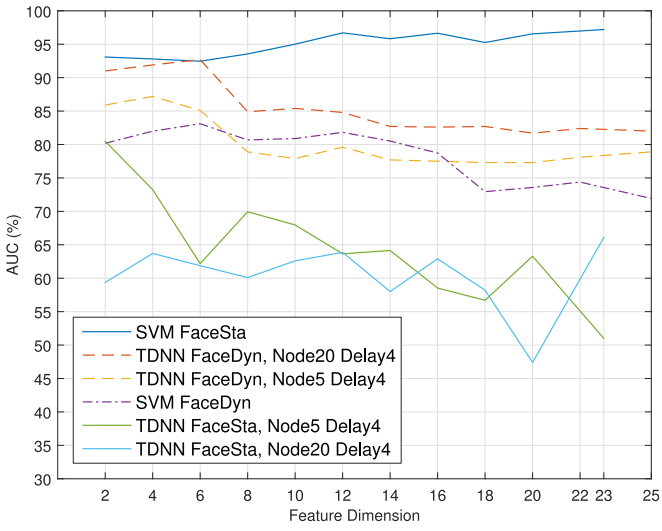

Fig. 5. AUC performances of laughter detection using discriminative static and dynamic facial features at varying dimensions.

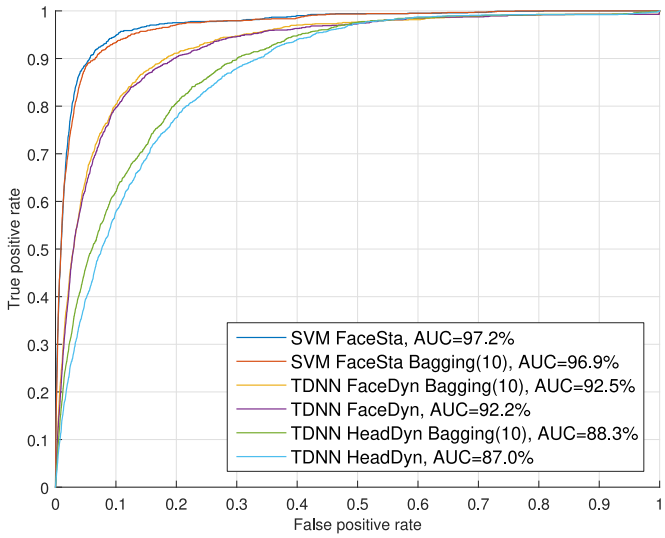

Fig. 6. ROC curves and AUC performances of laughter detection from head and facial features using different classifier-feature combinations with and without bagging.

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 in the upcoming experiments, we set the TDNN classifier with the dynamic facial features Δf^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 τ_d and n_h. Table 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 τ_d, we set the parameters as $\tau_d = 4$ and $n_h = 20$, which yield lower computational complexity due to a smaller delay parameter.

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. Fig. 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 percent AUC values. On the other hand, bagging does not help the SVM classifier with static facial features, which attains 97.2 percent AUC performance without bagging.

5.3 Fusion Results

The best performing classifier-feature combinations are integrated using the product rule defined in Section 4.5.2.

TABLE 3
The AUC Performances (%) of TDNN Classifiers with Dynamic Facial Features Δf^P at Feature Dimension 4 with Varying Delay τ_d and Number of Hidden Nodes n_h

	Hidden Nodes (n_h)		
Delay (τ_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

Four classifier-feature combinations, SVM-f^P, TDNN-Δf^P, TDNN-Δf^H and TDNN-f^A, are cumulatively populated, where all except SVM-f^P are with 10-fold bagging. Fig. 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 percent AUC performance. The best fusion scheme for three classifiers (Fusion3) is between SVM-f^P (FaceSta), TDNN-Δf^P (FaceDyn) and TDNN-f^A (Audio) with 98.3 percent AUC performance. Finally, the fusion of all four classifiers (Fusion4) attains 98.0 percent AUC performance.

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 percent false positive rate (FPR) as the anchor point on the ROC curve. At this anchor point, we first evaluate the

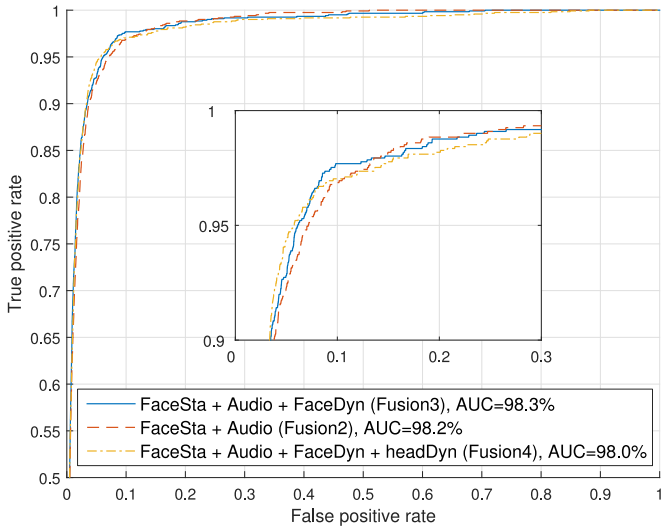

Fig. 7. ROC curves and AUC performances of laughter detection with fusion of best classifier-feature combinations, face static (SVM-f^P), face dynamic (TDNN-Δf^P), head dynamic (TDNN-Δf^H) and audio (TDNN-f^A)

TABLE 4
Recall Performances Under All, Clean, Cross-Talk and Noisy Laughter Conditions for Unimodal and Multimodal Classifiers at 2 Percent 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 5
Recall, Precision and F_1 Scores of Laughter Detection with Unimodal and Multimodal Classifiers at 2 Percent 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

recall performance under clean, cross-talk and noisy laughter conditions in Table 4. We observe that 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, which makes the fusion of four classifiers, Fusion4, to be the most robust one against noise.

Table 5 presents recall, precision and F_1 score performances of the unimodal and multimodal classifiers at 2 percent 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.

5.4 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 4.4 reveals that the facial laughter signal has a steady component, such as the contractions on the cheek region, which can be characterized with 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.

The best classifier according to AUC performance is the Fusion3, which is 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 percent AUC performance, it does not improve the fusion of all classifiers, i.e., the Fusion4 classifier. However, it brings 2 percent improvement in the recall rate of noisy laughter segments, as observed in Table 4. Hence head motion becomes valuable as a modality for laughter detection, especially in the presence of environmental acoustic noise.

Our multimodal laughter detection scheme benefits from the discriminative facial feature analysis presented in Section 4.4 in two aspects. First, as observed in Fig. 5, the AUC performance drops more than 10 percent 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 avails the possibility of real-time implementations. In fact, the SVM and TDNN classifiers that we employ are both 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 for analysis of engagement in HCI [46].

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 that we use in the training probably restricts the TDNN from better learning the non-laughter class. This could be the reason of the higher improvements that we obtain with bagging in the case of TDNN.

Table 6 positions our work in the context of the related work, and compares our approach with the state of the art in laughter detection and recognition. Existing work can be characterized as either recognition or detection frameworks. Recognition frameworks assume that the data has been presegmented into individual chunks, and the task reduces to classifying each chunk using standard classification algorithms. However, the data almost never comes in such presegmented 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

TABLE 6
Summary of the State of the Art in Laughter Detection and Recognition

Reference	Year	Task	Modality	Classifier	Database	Database Size	Performance
[19]	2005	Detection	Audio, Face (Image Based)	GMM	Own	3 clip, each 4-8 mins	Recall: 71%, Precision: 74
[18]	2008	Recognition	Audio, Head and Face Video Tracking	NN	AMI	Laughter: 58.4 sec, Speech: 118.1 sec	F1: 86.5%, ROC-AUC: 97.8%
[10]	2008	Detection	Audio, Face Video Tracking	GMM, HMM, SVM	AMI	Total: 25 mins (59% speech)	ROC-AUC: 93%
[17]	2009	Detection	Audio, Face Video Tracking	Gentle Adaboost	New York Times	Total: 72 mins	Accuracy: 0.77, Sensitivity: 0.65, Specificity: 0.77
[12]	2011	Recognition	Audio, Face Video Tracking	Neural Nets	AMI, SAL, DD, AVLC	Laughter: 33.6 min, Speech: 18.9 min	Average F1: 74.5
[22]	2012	Detection	Audio, Face and Body activity from video	HMM, SVM-GMM, ESN	FreeTalk	Total: 180 mins, Laughter: 289 sec, Speech-laugh: 307 sec	ESN Model, F1: 63%
[11]	2013	Detection	Audio, Face Video Tracking	TDNN	SEMAINE	Training: 77.3 min, Validation: 60.2 min, Test: 51.3 min	Recall: 41.9%, Precision: 10.3, F1: 16.4
[14]	2013	Recognition	Body Motion Capture	k-NN, RR, SVR, KSVR, KRR, MLP, RF, IR	Own	508 laughter, 41 non-laughter segments	RF Model, MSE: 0:011, CS: 0:91, TMR: 0:67, RL: 0:27, F1=74.4%
[6]	2015	Recognition	Body Motion Capture	k-NN, RR, SVR, KSVR, KRR, LASSO, MLP, MLP-ARD, RF, IR	UCL body laughter dataset	112 laughter, 14 non-laughter instances	RF Model, F-score: 0.60 (laughter class)
[5]	2016	Recognition	Body Motion Capture	SVM, RF, k-NN, NB, LR	MMLI	Laughter: 27 min 3 sec, Other: 46 min 18 sec	RF Model, Recall: 0.67, Precision: 0.66, F-score: 0.66
Our Method	2016	Detection	Audio, Face and Head Motion Capture	SVM, TDNN	IEMOCAP	Total: approx. 8 hours, Laughter: 382 sec	ROC-AUC: 98.3% (Fusion3)

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.

There are three frameworks that use only body movements [5], [6], [14] for performing recognition on pre-segmented laughters. Using various classification frameworks and different datasets, all these three studies show that body movements convey valuable information for laughter recognition. The performance figures reported by these methods are comparable to the ones which have audio and face modalities.

The first seven studies in the table [10], [11], [12], [17], [18], [19], [22], use audio and face information in recognition and detection tasks. Two of them [12], [18] used pre-segmented laughters. In [18], although they perform 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 leads to better performance compared to audio, and the head movement information has the lowest performance. The work in [12] 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 [10], [11], [17], [19], [22]. In [10], 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 percent AUC-ROC for audio and face fusion.

In [19], 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 percent of the whole data, which is quite high compared to ours (1.33 percent). This imbalance makes our task much more challenging.

Three threads of work [11], [17], [22] present proper detection algorithms on continuous data streams using relatively larger datasets. The method presented in [17] 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 [22] 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 [11] 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 percent 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 5 as 38.6 percent, while they have reported audio-facial scheme F_1 score as 16.4 percent.

6 Conclusion

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.

Acknowledgments

This work is supported by ERA-Net CHIST-ERA under the JOKER project and Turkish Scientific and Technical Research Council (TUBITAK) under grant number 113E324.

References

[1] L. Devillers, et al., "Multimodal data collection of human-robot humorous interactions in the joker project," in *Proc. Int. Conf. Affective Comput. Intell. Interaction*, Sep. 2015, pp. 348–354.
[2] W. Ruch and P. Ekman, "The expressive pattern of laughter," in *Emotion Qualia, and Consciousness*, Singapore: Word Scientific Publisher, 2001, pp. 426–443.
[3] S. Petridis and M. Pantic, "Audiovisual discrimination between speech and laughter: Why and when visual information might help," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 216–234, Apr. 2011.
[4] S. Cosentino, S. Sessa, and A. Takanishi, "Quantitative laughter detection, measurement, and classification—A critical survey," *IEEE Rev. Biomed. Eng.*, vol. 9, pp. 148–162, 2016.
[5] R. Niewiadomski, M. Mancini, G. Varni, G. Volpe, and A. Camurri, "Automated laughter detection from full-body movements," *IEEE Trans. Human-Mach. Syst.*, vol. 46, no. 1, pp. 113–123, Feb. 2016.
[6] H. J. Griffin, et al., "Perception and automatic recognition of laughter from whole-body motion: Continuous and categorical perspectives," *IEEE Trans. Affective Comput.*, vol. 6, no. 2, pp. 165–178, Apr. 2015.
[7] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
[8] K. P. Truong and D. A. Van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Commun.*, vol. 49, no. 2, pp. 144–158, 2007.
[9] K. Laskowski and T. Schultz, "Detection of laughter-in-interaction in multichannel close-talk microphone recordings of meetings," in *Machine Learning for Multimodal Interaction*. Berlin, Germany: Springer, 2008, pp. 149–160.
[10] B. Reuderink, M. Poel, K. Truong, R. Poppe, and M. Pantic, "Decision-level fusion for audio-visual laughter detection," in *Machine Learning for Multimodal Interaction*. Berlin, Germany: Springer, 2008, pp. 137–148.
[11] S. Petridis, M. Leveque, and M. Pantic, "Audiovisual detection of laughter in human-machine interaction," in *Proc. Humaine Assoc. Conf. Affective Comput. Intell. Interaction*, 2013, pp. 129–134.
[12] S. Petridis, M. Pantic, and J. F. Cohn, "Prediction-based classification for audiovisual discrimination between laughter and speech," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops*, Mar. 2011, pp. 619–626.
[13] K. R. Scherer, "Affect bursts," in *Emotions: Essays on Emotion Theory*. New York, NY, USA: Taylor & Francis, 1994.
[14] H. J. Griffin, et al., "Laughter type recognition from whole body motion," in *Proc. Humaine Assoc. Conf. Affective Comput. Intell. Interaction*, 2013, pp. 349–355.
[15] B. Schuller, F. Eyben, and G. Rigoll, "Static and dynamic modelling for the recognition of non-verbal vocalisations in conversational speech," in *Perception in Multimodal Dialogue Systems*. Berlin, Germany: Springer, 2008, pp. 99–110.
[16] D. Prylipko, B. Schuller, and A. Wendemuth, "Fine-tuning HMMs for nonverbal vocalizations in spontaneous speech: A multicorpus perspective," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Mar. 2012, pp. 4625–4628.
[17] S. Escalera, E. Puertas, P. Radeva, and O. Pujol, "Multi-modal laughter recognition in video conversations," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2009, pp. 110–115.
[18] S. Petridis and M. Pantic, "Fusion of audio and visual cues for laughter detection," in *Proc. Int. Conf. Content-Based Image Video Retrieval*, 2008, pp. 329–338.
[19] A. Ito, X. Wang, M. Suzuki, and S. Makino, "Smile and laughter recognition using speech processing and face recognition from conversation video," in *Proc. Int. Conf. Cyberworlds*, Nov. 2005, pp. 437–444.
[20] D. Cosker and J. Edge, "Laughing, crying, sneezing and yawning: Automatic voice driven animation of non-speech articulations," *Proc. Comput. Animation Social Agents*, 2009, pp. 21–24.
[21] E. Krumhuber and K. R. Scherer, "Affect bursts: Dynamic patterns of facial expression," *Emotion*, vol. 11, pp. 825–841, 2011.
[22] S. Scherer, M. Glodek, F. Schwenker, N. Campbell, and G. Palm, "Spotting laughter in natural multiparty conversations: A comparison of automatic online and offline approaches using audiovisual data," *ACM Trans. Interactive Intell. Syst.*, vol. 2, no. 1, pp. 4:1–4:31, Mar. 2012.

[23] J. Carletta, et al., "The AMI meeting corpus: A pre-announcement," in *Proc. 2nd Int. Conf. Mach. Learn. Multimodal Interaction*, 2006, pp. 28–39.
[24] B. B. Türker, S. Marzban, E. Erzin, Y. Yemez, and T. M. Sezgin, "Affect burst recognition using multi-modal cues," in *Proc. 22nd Signal Process. Commun. Appl. Conf.*, 2014, pp. 1608–1611.
[25] M. T. Suarez, J. Cu, and M. S. Maria, "Building a multimodal laughter database for emotion recognition," in *Proc. 8th Int. Conf. Language Resources Eval.*, May 2012, pp. 2347–2350.
[26] S. Petridis, B. Martinez, and M. Pantic, "The MAHNOB laughter database," *Image Vis. Comput.*, vol. 31, no. 2, pp. 186–202, Feb. 2013.
[27] G. McKeown, R. Cowie, W. Curran, W. Ruch, and E. Douglas-Cowie, "ILHAIRE laughter database," in *Proc. ES3 4th Int. Workshop Corpora Res. Emotion Sentiment Social Signals 8th Int. Conf. Language Resources Eval.*, 2012, pp. 32–35.
[28] R. Niewiadomski, M. Mancini, T. Baur, G. Varni, H. Griffin, and M. S. H. Aung, "MMLI: Multimodal multiperson corpus of laughter in interaction," in *Proc. Int. Workshop Human Behavior Understanding*, 2013, pp. 184–195.
[29] C. Busso, et al., "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources Eval.*, vol. 42, no. 4, pp. 335–359, 2008.
[30] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Commun.*, vol. 4, no. 1/2, pp. 33–60, 2003.
[31] C. Busso and S. Narayanan, "Recording audio-visual emotional databases from actors: A closer look," in *Proc. 2nd Int. Workshop Emotion: Corpora Res. Emotion Affect Int. Conf. Language Resources Eval.*, 2008, pp. 17–22.
[32] G. A. Bryant and C. A. Aktipis, "The animal nature of spontaneous human laughter," *Evol. Human Behavior*, vol. 35, no. 4, pp. 327–335, 2014.
[33] E. Bozkurt, E. Erzin, and Y. Yemez, "Affect-expressive hand gestures synthesis and animation," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2015, pp. 1–6.
[34] E. Bozkurt, E. Erzin, and Y. Yemez, "Multimodal analysis of speech and arm motion for prosody-driven synthesis of beat gestures," *Speech Commun.*, vol. 85, pp. 29–42, Dec. 2016.
[35] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoustical Soc. America*, vol. 111, no. 4, pp. 1917–1930, 2002.
[36] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image Vis. Comput.*, vol. 31, no. 2, pp. 137–152, 2013.
[37] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[38] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.
[39] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.
[40] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 37, no. 3, pp. 328–339, Mar. 1989.
[41] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.)*, vol. 42, no. 4, pp. 463–484, Jul. 2012.
[42] R. Akbani, S. Kwek, and N. Japkowicz, *Applying Support Vector Machines to Imbalanced Datasets*, Berlin, Germany: Springer, 2004, pp. 39–50.
[43] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.
[44] E. Erzin, Y. Yemez, and A. M. Tekalp, "Multimodal speaker identification using an adaptive classifier cascade based on modality reliability," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 840–852, Oct. 2005.
[45] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.
[46] B. B. Türker, Z Buçinca, E. Erzin, Y. Yemez, and M. T. Sezgin, "Analysis of engagement and user experience with a laughter responsive social robot," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 844–848.

Bekir Berker Turker received the BSc degree in electronics and communication Engineering from Izmir Institute of Technology. He is currently working toward the PhD degree in the Electrical Engineering Department, Koc University. His research interests include human-computer interaction, affective computing, social robotics, and audio-visual signal processing.

Yucel Yemez received the BS degree from Middle East Technical University, Ankara, in 1989 and the MS and PhD degrees from Bogazici University, Istanbul, respectively, in 1992 and 1997, all in electrical engineering. From 1997 to 2000, he was a postdoctoral researcher in the Image and Signal Processing Department of Telecom Paris (ENST). Currently, he is an associate professor in the Computer Engineering Department, Koc University, Istanbul. His research interests include various fields of computer vision and graphics.

T. Metin Sezgin received the graduate (summa cum laude) degree with Honors from Syracuse University, in 1999. He received the MS degree from the Artificial Intelligence Laboratory, Massachusetts Institute of Technology, in 2001. He received the PhD degree in 2006 from Massachusetts Institute of Technology. He subsequently moved to University of Cambridge, and joined the Rainbow Group at the University of Cambridge Computer Laboratory as a postdoctoral research associate. He is currently an associate professor in the College of Engineering, Ko University, Istanbul. His research interests include intelligent human-computer interfaces, multimodal sensor fusion, and HCI applications of machine learning. He is particularly interested in applications of these technologies in building intelligent pen-based interfaces. His research has been supported by international and national grants including grants from DARPA (USA), and Turk Telekom. He is a recipient of the Career Award of the Scientific and Technological Research Council of Turkey.

Engin Erzin (S'88-M'96-SM'06) received the BSc, MSc, and PhD degrees from Bilkent University, Ankara, Turkey, in 1990, 1992, and 1995, respectively, all in electrical engineering. During 1995-1996, he was a postdoctoral fellow in the Signal Compression Laboratory, University of California, Santa Barbara. He joined Lucent Technologies in September 1996, and he was with the Consumer Products for one year as a member of Technical Staff of the Global Wireless Products Group. From 1997 to 2001, he was with the Speech and Audio Technology Group of the Network Wireless Systems. Since January 2001, he is with the Electrical & Electronics Engineering and Computer Engineering Departments, Koç University, Istanbul, Turkey. His research interests include speech signal processing, audio-visual signal processing, human-computer interaction, and pattern recognition. He has served as an associate editor of the *IEEE Transactions on Audio, Speech & Language Processing* (2010-2014) and as a member in the IEEE Signal Processing Education Technical Committee (2005-2009). He was elected as the chair of the IEEE Turkey Section in 2008-2009. He is a senior member of the IEEE.