Multimodal Sensing and Data Processing for Speaker and Emotion Recognition using Deep Learning Models with Audio, Video and Biomedical Sensors

by

Farnaz Abtahi

A dissertation submitted to the Graduate Faculty in Computer Science in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

February 2018

This manuscript has been read and accepted for the Graduate Faculty in Computer Science in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Date	Professor Zhigang Zhu Chair of Examining Committee
Date	Professor Robert Haralick Executive Officer

Professor Zhigang Zhu (City College of New York and CUNY Graduate Center)

Professor Tony Ro (CUNY Graduate Center)

Professor Yingli Tian (City College of New York)

Professor Lijun Yin (SUNY Binghamton)

Supervisory Committee

THE CITY UNIVERSITY OF NEW YORK

Abstract

Multimodal Sensing and Data Processing for Speaker and Emotion Recognition using Deep Learning Models with Audio, Video and Biomedical Sensors

by

Farnaz Abtahi

Advisor: Professor Zhigang Zhu

The focus of the thesis is on Deep Learning methods and their applications on multimodal data, with a potential to explore the associations between modalities and replace missing and corrupt ones if necessary. We have chosen two important real-world applications that need to deal with multimodal data: 1) Speaker recognition and identification; 2) Facial expression recognition and emotion detection.

The first part of our work assesses the effectiveness of speech-related sensory data modalities and their combinations in speaker recognition using deep learning models. First, the role of electromyography (EMG) is highlighted as a unique biometric sensor in improving audio-visual speaker recognition or as a substitute in noisy or poorly-lit environments. Secondly, the effectiveness of deep learning is empirically confirmed through its higher robustness to all types of features in comparison to a number of commonly used baseline classifiers. Not only do deep models outperform the baseline methods, their power increases when they integrate multiple modalities, as different modalities contain information on different aspects of the data, especially between EMG and audio. Interestingly, our deep learning approach is word-independent. Plus,

the EMG, audio, and visual parts of the samples from each speaker do not need to match. This increases the flexibility of our method in using multimodal data, particularly if one or more modalities are missing. With a dataset of 23 individuals speaking 22 words five times, we show that EMG can replace the audio/visual modalities, and when combined, significantly improve the accuracy of speaker recognition.

The second part describes a study on automated emotion recognition using four different modalities – audio, video, electromyography (EMG), and electroencephalography (EEG). We collected a dataset by recording the 4 modalities as 12 human subjects expressed six different emotions or maintained a neutral expression. Three different aspects of emotion recognition were investigated: model selection, feature selection, and data selection. Both generative models (DBNs) and discriminative models (LSTMs) were applied to the four modalities, and from these analyses we conclude that LSTM is better for audio and video together with their corresponding sophisticated feature extractors (MFCC and CNN), whereas DBN is better for both EMG and EEG. By examining these signals at different stages (pre-speech, during-speech, and post-speech) of the current and following trials, we have found that the most effective stages for emotion recognition from EEG occur after the emotion has been expressed, suggesting that the neural signals conveying an emotion are long-lasting.

Acknowledgements

During my PhD research, I enjoyed a multi-faceted collaboration environment, both cross-disciplines and between academia and industry.

The major part of my work was a collaboration between The City College Visual Computing Research Laboratory (CCVCL), at the Department of Computer Science of the City College of New York, and the Ro Lab, at the Department of Psychology and Biology at CUNY Graduate Center. In particular, the data that we collected for both speaker recognition and facial expression and emotion processing was collected and processed using equipment from both labs. Machine Learning, Computer Vision, Image Processing and other Signal Processing techniques were used to extract information from visual and audio data (by help of Prof. Zhigang Zhu and my fellow lab members at CCVCL), while Neuroscience and Psychology skills, sensors and devices from the Ro Lab (Prof. Tony Ro and his students) were used to collect and process biomedical data including EMG and EEG signals. The data for my research was collected with the help of Ms. Olesya Medvedeva, who worked with me as the lab assistant during her participation in CCNY REM program in 2015.

For facial expression recognition and feature extraction from videos and images, we collaborated with a fellow lab member, Dr. Wei Li. Apart from contributing to the early stages of his work, the developed ROI-Net (Li, Abtahi & Zhu, 2017) has been used as a feature extractor in Chapter 3. We also participated in Zahn Innovation Center's entrepreneurship competition as the EmoTrain team, consisting of Wei Li, Christina Tsangouri and Farnaz Abtahi, at the City College of New York, advised by Ms. Celina Cavalluzzi, the director of Goodwill Day Services. The goal of the team was to develop mobile and web educational applications for autistic

individuals to assist them with development of emotion recognition and emotion expression skills. EmoTrain finished the competition as a semifinalist.

Another collaboration took place during summer research internships in summers of 2013 and 2014 at Palo Alto Research Center East (PARC East), formerly Xerox Research Center Webster (XRCW, 2013) under supervisory of Mr. Aaron Burry. During the two internships at this research center, existing deep learning algorithms using DBNs were explored for license plate character segmentation and recognition in Xerox's Automatic License Plate Recognition (ALPR) System. This led to development of a patented new character segmentation algorithm that combines Reinforcement Learning techniques and Deep Learning methods. While this is not included in the thesis, the DBN models are common in both scenarios; more details on the background of the algorithm can be found in (Abtahi & Fasel, 2011). The character segmentation method and results are presented in the patent (Burry & Abtahi, 2015) and the conference paper (Abtahi, Zhu & Burry, 2015); both are included in the List of Candidate's Publication section.

I would like to express my sincere gratitude to my advisor, Professor Zhigang Zhu, for his endless support, limitless patience, motivation, and knowledge. His guidance helped me in overcoming numerous obstacles I have been facing through my research, and my academic and personal life in general, over the past five years. I would also like to thank my co-advisor, Professor Tony Ro, for his continuous support, encouragement, and immense knowledge that he generously shared with me throughout my PhD research. I could not have imagined having better mentors for my PhD study.

Besides my advisor and co-advisor, I would like to thank the rest of the committee members: Professor YingLi Tian, who, aside from her constructive feedbacks on my research, has always been an example of a strong, fearless and inspiring woman to me; Professor Lijun Yin, who

despite our short acquaintance, kindly agreed to support me in multiple ways, through his insightful comments on my work during his visit to our lab at CCNY, and by accepting to be in my PhD committee; and finally, Mr. Aaron Burry, who has always been supporting me openheartedly, since my first internship at Xerox in 2013.

My sincere thanks also goes to my fellow lab members, Wei Li, Greg Olmschenk, Hao Tang, Edgardo Molina, Wai Khoo, Feng Hu, Martin Goldberg and Christina Tsangouri, for their feedback, cooperation, support and of course friendship, that will hopefully last a lifetime.

Last but not least, I would like to thank my family, without whom I would not have been able to come this far, especially my mother, father and sister, for all their encouragements and confidence boosts.

This research was supported by NSF EFRI Award #1137172, NSF BCS Awards #1358893 and #1561518, and an Enhanced Chancellors Fellowship from the CUNY Graduate Center.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

In most real-world applications, dealing with multimodal data is inevitable due to the nature of the task. This requires machine learning methods that are capable of efficiently combining their learned knowledge from multiple modalities. In traditional machine learning methods such as Support Vector Machines (SVMs) (Cortes, 1995), learning is performed by training a separate SVM on each individual modality and combining the results by voting, weighted average or other probabilistic methods.

A very important aspect of multimodal learning that is missed in these approaches is the ability to automatically *learn* the features of various modalities, to effectively *integrate* multiple modalities, and to inherently *associate* different modalities. All these can be easily achieved by utilizing deep learning methods, as they are capable of extracting task-specific features from the data and learning the relationship between modalities through a shared representation. This shared representation of the data which reveals the association between different modalities makes the trained structure a generative model. That is, the model would be able to make the best use of the complementary information of different modalities, and handle missing

modalities as long as the relationship between the absent modality is efficiently learned by the model.

The importance of dealing with missing modalities are two-folds. First a modality is available during the training phase but might be missing or corrupted during the online recognition phase. Second, a certain modality may not be useful for real applications but it is advantageous to train a more robust model with this modality together with others that would be always available. While the thesis will work on both deep-learning-based multimodal feature extraction and integration, an interesting direction for future work is to focus on missing modalities.

In this thesis, our main focus is on the application of Deep Learning models on multimodal data and their capabilities to learn the association between modalities. As mentioned above, this learned association can potentially enable the model to deal with missing or corrupt modalities. We have chosen two important real-world applications with multimodal data: 1) Speaker recognition and identification, and 2) Facial expression recognition and emotion detection.

The first application assesses the effectiveness of speech-related sensory data modalities, including voice, mouth movements from images, and biometric information reflecting facial muscle movements, and their combinations in speaker recognition using deep learning models. We first highlight the role of a unique biometric sensory input captured through electromyography (EMG) and show that it improves the accuracy of audio-visual speaker recognition in combination with other modalities or as a substitute in noisy or poorly-lit environments. Secondly, we confirm the effectiveness of deep learning on multimodal data through empirical analyses and show that deep learning models have higher robustness to all types of features in comparison to a number of commonly used baseline classifiers. Not only do

deep models outperform the baseline methods, they can successfully integrate multiple modalities and further increase the accuracy, as different modalities contain information on different aspects of the data, particularly among EMG and audio.

Our analyses prove a number of interesting hypotheses, such as independence of different modalities, as long as they belong to the same individual. In other words, we can match and combine modalities from different utterances by the same person and still achieve high accuracy in speaker identification. We also show that EMG can potentially replace voice in the tasks where voice is unavailable, corrupt or noisy, as it can achieve comparable accuracy as voice in speaker identification. This part of our work is thoroughly studies and explained in Chapter 2.

Chapter 3 describes a study on automated emotion recognition using four different modalities – audio, video, electromyography (EMG), and electroencephalography (EEG). We have collected a valuable dataset comprising the above four modalities recorded during simultaneous speech and facial expression. This dataset will be published for research purposes.

Our goal in this part of our work is not to combine the modalities for facial expression and emotion recognition, but rather to focus on the unique characteristics of different modalities and their role in understanding the trend of emotional states over time. We study this by examining these signals at different stages (pre-speech, during-speech, and post-speech) of the current and following trials. In addition, we highlight three different aspects of emotion recognition: 1) model selection, 2) feature selection, and 3) data selection. Both generative models (such as Deep Belief Networks) and discriminative models (such as Long-Short Term Memory Networks) were applied to the four modalities. The analyses show that each of these categories of models is useful for certain modalities. This leads us to another future direction, where we would combine generative and discriminative models for multimodal emotion analysis.

1.2 Contributions

Several aspects of our work are useful for proving insights and theoretical hypotheses or/and being used in practical systems and real-world applications. Here is a summary of the contributions of our research:

- ***Sensors and datasets.*** We have explored different types of sensors, each capturing one or more of the modalities used in our research. In other words, we have performed multimodal sensing and data processing, including audio, video and biomedical sensors, particularly with novel sensors such as EMG/EEG for obtaining biometric signatures. Two datasets have been collected, one for the speaker recognition task and the other one for facial expression recognition. These datasets can be extended or used in similar applications as-is.
- ***Deep learning models.*** We studied several multimodal deep learning models in detail, used CNNs to learn image features for facial expression recognition, implemented a multimodal DBN and applied it to the speaker recognition data that we collected. Both generative models (DBNs) and discriminative models (LSTMs) were applied to four sensor modalities of the facial expressions dataset as well..
- ***Learned features.*** The DBN and CNN can work as a feature extractor, meaning that the hidden units are able to dig deeper into the data and extract features that are not visible to conventional feature extraction methods. These features have great potential in transfer learning, which is build based on the intuition that generalization may occur not only within tasks, but also across tasks. In particular, a pre-trained CNN that has been successfully used for facial expression recognition.
- ***Association of modalities.*** We empirically prove that having an extra modality during training is effective in improving the recognition rate and also capturing the association

between modalities in case the model needs to be used generatively. Training the model on an extra modality would leverage the recognition power of the model. This has been done in the speaker identification task.

- ***Analysis of different modalities.*** In emotion recognition, both generative models (DBNs) and discriminative models (LSTMs) were applied to the four modalities, and from these analyses we conclude that LSTM is better for audio and video together with their corresponding sophisticated feature extractors (MFCC and CNN), whereas DBN is better for both EMG and EEG. By examining these signals at different stages (pre-speech, during-speech, and post-speech) of the current and following trials, we have found that the most effective stages for emotion recognition from EEG occur after the emotion has been expressed, suggesting that the neural signals conveying an emotion are long-lasting.

Chapter 2

Multimodal Speaker Recognition

2.1 Introduction

Automatic speaker recognition (including identification and verification) is the process of automatically recognizing the speaker based on the information included in speech (Furui, 1997). Speaker identification is the process of determining which speaker provides a given utterance. Speaker verification, on the other hand, is the process of accepting or rejecting the identity claim of a speaker. Most applications in which a voice is used as the key to confirm the identity of a speaker are classified as speaker verification. These applications include but are not limited to voice dialing, telephone banking, telephone shopping, database access, information services, voicemail, remote access to computers, security control, etc.

The majority of existing speaker recognition systems are on the basis of the speaker's voice. In these systems, the task of speaker recognition consists of two major phases (Gang and Hansen, 2014): enrollment (data collection and/or training) and verification and/or identification (test). The enrollment phase deals with collecting the data by recording the speaker's voice and typically extracting features suitable for voice processing. During the verification/identification phase, the speaker's voice is compared against the existing voices or extracted features that were previously collected and processed during the enrollment phase.

There are two categories of speaker recognition systems (Reynolds, 2002): text-dependent and text-independent. In text-dependent systems, the words used during enrollment is the same as the ones used for testing the system, while in text-independent systems, the words used during the two phases could be different. Basically text-independent speaker recognition systems are independent of what the speaker says. This makes these systems suitable for identification purposes, as the speakers have to be identifiable no matter what they say. Due to the fact that the words used during test is not necessarily the same as the words used for enrollment, text-independent systems are more flexible and can also apply speech recognition to determine what the speaker is saying.

Although the most informative if captured with acceptable quality and clarity, using voice as the sole modality or source of information makes speaker recognition equivalent to voice recognition. However, relying on voice for speaker recognition would be difficult or sometimes impossible in noisy environments or in applications where voice is completely unavailable due to sensor defectiveness or security reasons. Other information modalities, on the other hand, which define a person's speaking style and accent, can also be extremely valuable for these types of tasks. If visual information such as images or videos is available during speech, it can be very useful as it includes lip movements and also facial expressions that are unique and specific to the person or an ethnic group or nationality, in addition to identity information from the face.

Another type of information that has not yet received enough attention as a biometric for speaker recognition is electrical activity of facial muscles, or electromyography (EMG). These signals, which are also called myoelectric signals (MES), represent facial muscle movements during a facial activity. Several factors affect EMG, including body mass, muscle fiber pattern, muscle size, motion of subject, neuromuscular activity, neurotransmitter activity in different

areas within the muscle, different density of bone, changes in blood flow in the muscle, fatigue, skin conductivity, motor unit firing pattern, motor unit paths, distribution of heat in the muscle, skin- fat layer, motor unit recruitment order and characteristics of muscle, strength and force generated by the muscle (Suresh et al., 2014). The combination of these factors is very distinct and unique to each person and might be effectively used to determine speaking style and identity. The information conveyed by EMG would be especially useful when the environment is noisy or/and poorly illuminated. In those situations, relying on voice or visual data might not result in accurate recognition of speakers. Furthermore, the study of the connection between facial muscle movements and audio/video data of a speaker has more scientific value in finding the association between those modalities, as well as capturing the style of speaking, which is unique to each individual, for applications such as forensics and digital system access. Another important benefit of incorporating EMG in voice or image-based speaker recognition systems is that the EMG signals cover a unique area of information space that audio or visual data may miss. The speakers that are recognized incorrectly using their voice or image, might be recognizable using their EMG activity and vice versa. Hence, these modalities are complementary.

Typical EMG systems used in labs are not 100% portable. However, with the use of recent highly portable sensors (BioSemi, Delsys, SparkFun), EMG-based systems can provide a very promising alternative to audio and audio-visual approaches for speaker recognition. The facial muscles that are most involved in human speech activities and the advantages of using facial EMG signals to infer states have been reported in the literature (Van Boxtel, 2010). These signals can be much more reliable than visual information since they can capture the tiniest facial muscle activities that cannot be detected by the human eye.

In this chapter, we compare and combine multiple modalities, namely voice signals, visual lip motion data, and EMG signals for speaker recognition. An ideal model for combining these speech-related modalities has to be capable of successfully combining them in different ways, capture their associations, and be robust to noise and unsynchronized data from different modalities. The model that we use in this chapter is a multimodal Deep Belief Network (DBN).

Deep learning in general and specifically DBNs have proved their power in several applications, both individually and as part of a multimodal model (Ngiam et al., 2011; Srivastava and Salakhutdinov, 2012). DBNs work very well when the data is one-dimensional. More importantly, DBNs are generative models, meaning that they are able to run backwards and generate samples similar to the ones included in the data presented to the model so far. In a multimodal setting, using a DBN to learn a shared representation of the data will enable us to determine the relationship between different modalities. This is particularly useful in applications where one modality or a part of it is missing or corrupt. In speaker recognition for instance, this would play a huge role, as EMG and voice are prone to being missing or too noisy to be of any value. For EMG, this could be due to non-portable EMG sensors and long preparation time that makes collecting EMG data impractical when the system is deployed. On the other hand, for audio signals, this issue occurs when the environment is too noisy, the voice is not captured either because the device has not worked properly, or the audio should not be transmitted (e.g., for security reasons). The multimodal DBN would not only be able to deal with missing modalities, it has the ability to estimate the value of missing parts of the data.

The idea is that during training of the model, all modalities are available and the joint distribution of the training data is learned successfully. During test though, either EMG or voice could be missing, but the model is still able to recognize the speaker using the available

modalities, given the learned joint distribution of the entire training data. In this chapter, we show that the accuracy of speaker recognition using EMG is comparable with the accuracy using voice, which leads to the conclusion that these two modalities can complete or substitute one another if need be.

To justify the use of DBNs for multimodal speaker recognition, we also show their superiority on speaker recognition over other common classification models used in similar applications such as Gaussian Mixture Models (GMM) (Reynolds, 2015), i-vector features, also called total variability space approach (Dehak et al., 2011), in combination with Probabilistic Linear Discriminant Analysis (PLDA) (Ioffe, 2006), as well as Support Vector Machines (SVMs) (Cortes and Vapnik, 1995; Hosseini et al., 2014) and Random Forest (RF) classifiers (Ho, 1995).

Due to the capabilities of GMM to model arbitrary densities, these models have been vastly used in speech as well as speaker recognition. In such applications, Gaussian components are used to represent speaker-dependent spectral shapes (Zeng and Li, 2011).

PLDA-based i-vector speaker recognition systems have been very successful in speaker recognition and are considered state-of-the-art in this field. I-vector representation consists of mapping a sequence of frames for a given utterance into a low-dimensional vector space, referred to as the total variability space, based on a factor analysis technique (Dehak et al., 2011). The advantages, disadvantages and possible improvements to i-vector-based speaker recognition and verification are thoroughly explored in (Kanagasundaram, 2014).

SVM-based speaker recognition and verification have also been studied in literature (Kanagasundaram, 2014). SVMs are proved to be effective when the classes are not linearly separable in a classification problem. In addition, SVMs work well if the problem has a high

dimensional space. Both of these assumptions hold for the problem with which we are dealing. On the other hand, the main advantage of RFs (and tree ensembles in general) is that, because of the way they are constructed using bagging (Breiman, 1996) or boosting (Schapire, 1990), these methods can very well handle high dimensional spaces as well as large numbers of training examples. Compared to other tree ensemble algorithms such as Gradient Boosted Decision Trees (GBDTs) (Friedman, 2001), RFs have fewer hyperparameters to tune and are also less prone to overfitting. They can almost work "out of the box" and this has made them very popular.

To the best of our knowledge, no multimodal model that includes EMG exists for speaker recognition. The results of our experiments show that not only do the individual DBNs trained on single modalities generally outperform the baseline methods we used for comparison, the multimodal DBN is more powerful than unimodal DBNs and can achieve high accuracy in speaker recognition. Furthermore, we have performed various analyses to show the advantages of the EMG modality for speaker recognition: (1) The performance of speaker recognition using EMG only is comparable to the voice-only results. (2) EMG alone outperforms the lip-motion only approach, implying that facial muscle movements include more essential information of individuality than lip motion data. (3) We show that the EMG modality significantly improves the accuracy of voice-only speaker recognition by more than 9% when combined with voice, which is 2% higher than the accuracy of the combination of lip motion and voice. This implies that EMG provides additional information about the speakers that the lip data has missed. (4) Two analyses show that EMG captures more of the essentials of speech than the visual lip motion data and complementary information as compared to audio, which to our knowledge is the first study using multimodal data including EMG for speaker recognition. (5) We also show that the DBN-based approach for multimodal speaker recognition is both word-independent (we

use word-independent instead of text-independent since sensory data when a speaker was uttering a single word, not written text or utterances of sentences from text, were used for identification) and tolerant to unmatched/asynchronized multimodal data, meaning that the audio and EMG data do not have to correspond to the same spoken word as long as they belong to the same person, making this approach more flexible in real world applications.

2.2 Related work

2.2.1. Multimodal speaker recognition

Although the modalities involved in speaker recognition can be used separately, we expect their combination to be much more powerful in recognizing the speaker. A number of multimodal speaker recognition approaches have been proposed in the literature. For instance, a boosting-based approach has been proposed that combines audio and visual information (Zhang et al., 2008). The audio and visual information are fused at the feature level. Then boosting is used to select effective features. The proposed approach outperforms single modalities and is applied in real-time distributed meetings.

The method proposed in Hazen et al. (2007) is a face- and audio-based multimodal speaker detection approach for mobile devices. GMM and SVM are used for face and audio classifications, then a linear combination of the two models is proposed to detect speakers using the mobile device. The authors also employ the posterior union model and universal compensation to make the model more robust to corrupted features.

The multimodal model in Çetingül et al. (2006) combines lip motion, lip texture and audio for speaker/speech recognition. Mel Frequency Cepstral Coefficients (MFCCs) are used to extract features from audio data. The features from the lip texture are described with the 2D

Discrete Cosine Transform (2D-DCT) coefficients of the luminance components and the lip motion features are the dense motions of the lip region. The paper proposed to use the Reliability Weighted Summation (RWS) to perform the model fusion to make sure the fusion result is better than the single modalities.

The multimodal speaker verification approach described in Zhang and Broun (2001) uses lip features, in addition to audio information from their existing speaker verification system. Color and edge information are combined within a Markov random field (MRF) framework to localize the lips. A polynomial classifier is then applied to geometric features of the lips (including height and length of the lips and visibility of teeth and tongue) for person recognition. Finally, an integration approach based on a Bayesian model is used to combine the visual and audio modalities.

The multimodal model proposed in Roy and Shukla (2013) is also based on face and voice information. An Artificial Neural Network (ANN) is applied to the features extracted from face and voice. The text-independent speaker recognition method proposed in Nakagawa et al. (2004) combines a speaker-specific GMM with a syllable-based Hidden Markov Model (HMM).

A model-based feature extraction method which employs physiological characteristics of facial muscles producing lip movements is proposed in Asadpour et al. (2006). Their approach, which is intended for security systems, adopts the intrinsic properties of muscles such as viscosity, elasticity, and mass, which are extracted from the dynamic lip model, based on the assumption that these parameters are exclusively dependent on the neuro-muscular properties of the speaker and very hard to imitate. These parameters are applied to a HMM audio-visual identification system.

Ren et al. (2016) describe a novel multimodal Long Short-Term Memory (LSTM) architecture which seamlessly unifies both visual and auditory modalities from the beginning of each sequence input. The key idea is to extend the conventional LSTM by not only sharing weights across time steps, but also sharing weights across modalities. They show that modeling the temporal dependency across face and voice can significantly improve the robustness to content quality degradations and variations. They also found that the proposed multimodal LSTM is robustness to distractors, namely the non-speaking identities. The authors applied their multimodal LSTM to The Big Bang Theory dataset and showed that their system outperforms the state-of-the-art systems in speaker identification with lower false alarm rate and higher recognition accuracy.

2.2.2. Speaker and speech recognition using EMG data

Although EMG signals have not been used much for speaker recognition, they have been tested for speech recognition (Chan et al., 2006; Lee, 2008; Quan et al., 2009; Wand and Schultz, 2011), in which temporal information from the data is more critical than in speaker recognition.

In Quan et al. (2009), the authors used a phoneme-based approach that has the capacity in expanding the vocabulary of words without new training. The system utilizes the commonly used HMM in speech recognition and MFCCs as voice features. However, the segmentation of spoken words into phonemes would be a challenging issue. An earlier study by the same group (Chan et al., 2006) had used a plausibility method of combining voice and EMG results, based on the mathematical framework of evidence theory, and achieved improved classification performance with noisy signals in both audio and EMG channels.

Lee (2008) highlights the strong relationship between human voices and the movement of articulatory facial muscles and implements an automatic speech recognition scheme that uses

solely surface EMG signals. Three EMG electrodes are used in this work and their locations are determined heuristically, based on a trial-and error approach. They utilize a HMM to build a model for state observation density when multichannel observation sequences are given. A global control variable is introduced to reflect the dependencies between the EMG channels. The work presented in Wand and Schultz (2011) is done based on the fact that EMG can be used to create silent speech interfaces, since the EMG signal is available even when no audible signal is transmitted or captured. The authors present a session-independent system, which shows that a system trained on multiple recording sessions of one and the same speaker yields a reasonable performance and recognizes test data from unseen sessions more robustly than a similarly large recognizer trained on data from just one session. They also show that the increased robustness of a session-independent system helps to cope with the difference between normal and silently articulated speech. The ability of the system to cope with increasing vocabulary sizes is further tested on a vocabulary of more than 2000 words.

Evaluating the effectiveness of EMG in speaker identification barely appears in the current literature. One group of researchers has worked on EMG as the only modality for person identification (Suresh et al., 2011; Suresh et al., 2014). As the authors have thoroughly explained, EMG is a great biometric that is very distinct in each person, since it is affected by several factors as mentioned in the Introduction. Only a single EMG channel has been used in their experiments from a sensor on the flexor carpi ulnaris muscle. In their most recent work, 100 individuals were used in data collection as opposed to 49 in their previous work. The data was recorded in three sessions on different days. A Non-Uniform Filter Bank (NUFB) technique was used to extract features from the EMG signal and GMM was used to generate person models from NUFB features. The highest accuracy of 97.96% was achieved when their training slots

were from the same sessions as one of the sessions present in training set, e.g. if the model is trained on sessions 1 and 2 and tested on session 2. When the model was trained and tested on slots from different sessions, the highest achieved accuracy was 85.71 for training on sessions 1 and 3 and testing on session 2.

2.3 Multimodal deep learning model and baseline methods

Machine learning from multiple modalities using deep learning models has recently been gaining a lot of attention (Ngiam et al., 2011; Srivastava and Salakhutdinov, 2012). Multimodal models usually improve recognition rates and are robust to missing modalities. They also capture associations between heterogeneous modalities by learning mid-level feature representations.

We have chosen DBN to be the building block of our multimodal deep learning model because they perform well on one-dimensional data, where samples are 1xN vectors. Moreover, DBNs are generative models, which makes them capable of utilizing the learned associations between different modalities to run backward and generate samples similar to the real data. For completeness, before getting into the details of the multimodal DBN, we give a brief overview of the DBN and its characteristics.

2.3.1 The basics of deep belief networks

DBNs are probabilistic graphical models that are built by stacking up Restricted Boltzmann Machines (RBMs) (Abtahi and Fasel, 2011; Hinton et al., 2006). An RBM is an undirected graphical model that consists of one layer of visible and one layer of hidden Bernoulli units. There are no connections between units of the same layer, but the two layers are fully connected to each other. Connections between layers are bidirectional and symmetric, so the weights are

also shared between both layers. Figure 1 (left) shows an RBM with 4 visible and 3 hidden units. A sample DBN is illustrated in Figure 1 (right).

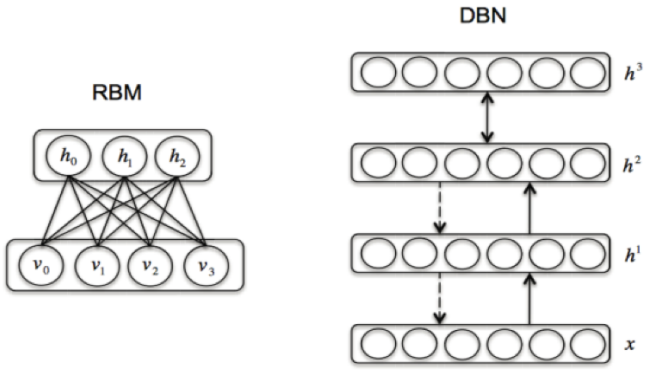

Figure 1. An RBM with 4 visible (input) and 3 hidden units (left) and a DBN with the same number of units in all layers (right) (Abtahi and Fasel, 2011)

The effect of pre-training is studied in detail in Erhan et al. (2009). They explain that the reason why pre-trained DBNs work much better than traditional neural networks is that pre-training initializes the parameters of the DBN in a more desirable area of parameter space where a better local optimum can be found. Therefore, pre-training introduces a bias towards configurations of the parameters that the supervised learning phase can explore.

2.3.2 Multimodal deep belief networks

Similar to all multimodal deep learning algorithms, multimodal DBNs can combine multiple input domains such as images, audio and speech, video, text, robotics sensors, time series data, etc. Generally, different modalities can be combined in two different ways. First, inputs from all modalities can be concatenated and used as a set of single input vectors in a regular DBN. This is usually not the best option because the information from different modalities are not of the same type, but the model will try to learn unimodal features form the concatenated input.

A better way of combining the modalities, which is also used in this chapter, is to train sub-DBNs on each modality and then add one or more layers to combine those models and learn a

shared representation of all modalities. An example of this type of multimodal DBN is shown in Figure 2. In this case, lip images, audio data, and EMG inputs are combined using the shared representation and classification occurs at the top two layers of the model.

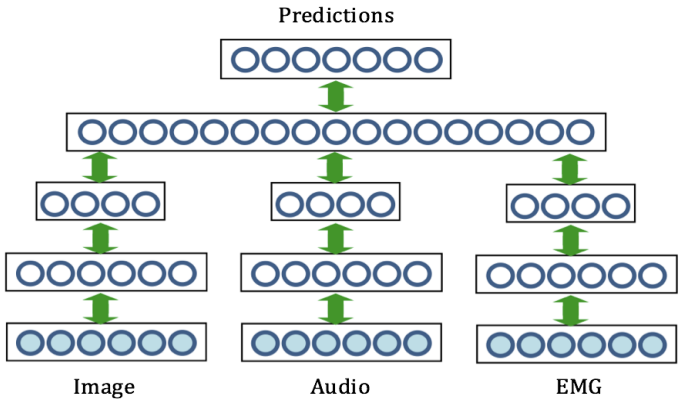

Figure 2. A multimodal deep learning model for combining image, audio and EMG

In our proposed multimodal DBN model, a single DBN is first trained on each modality in a generative way, i.e. for extracting features from the input. Then a single layer is added on top of all of the models to learn a shared representation of the outputs of the individual DBNs. The last layer is then added on top of the shared representation layer to make the entire model a classifier.

2.3.3 Baseline methods

The baseline methods used in this part of the work include Support Vector Machines (SVMs), Random Forest (RF), Gaussian Mixture Model (GMM) and i-Vector in combination with Probabilistic Linear Discriminant Analysis (PLDA). In this section, we briefly review these four methods.

- ***Support Vector Machines:***

SVMs were introduced in 1992 by Boser, Guyon and Vapnik (Cortes, 1995) and have shown good empirical performance in many applications (bioinformatics, text, image recognition, etc.). Figure 3 shows the basic idea of the SVM for the simple case of 2D 2-class classification

problem on a set of input-output pairs *{(x,y)}*. Assuming that we represent the input/output sets as $X = \{x\}$ and $Y = \{y\}$, where $y = +1$ or -1, the goal is to learn the function $y = f(x,a)$, where α are the parameters of the function. In the example of Figure 3, f can be defined as $f(x,\{w,b\})=sign(w \bullet x+b)$. So the goal is to find the best set of parameters w and b so that the margin between the two classes (shown with a 2-way green arrow) is maximized.

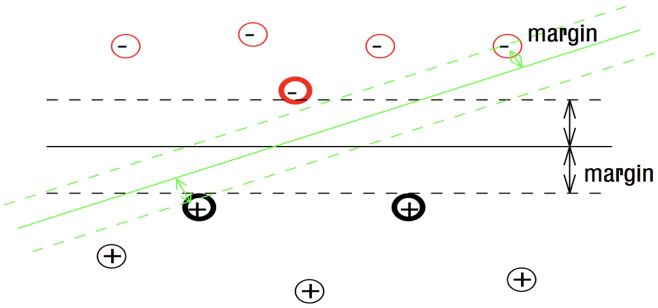

Figure 3. SVM on 2D 2-class data

For inseparable classes, the function f is nonlinear and hard to find. In this case, the trick is to map data into a richer feature space including nonlinear features and then construct a hyperplane in that space to separate the classes in a linear way. This is shown in Figure 4. Formally, we need to preprocess the data with $x \rightarrow \Phi(x)$, and then learn the map from $\Phi(x)$ to y: $f(x)=w \bullet \Phi(x)+b$.

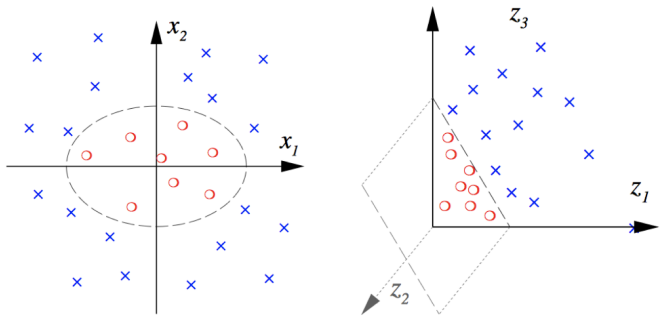

Figure 4. Mapping inseparable 2D data to a separable feature space

- ***Random Forests:***

During the recent decades, there has been a lot of interest in ensemble learning — methods that generate many classifiers and aggregate their results. Two well-known methods are boosting

(Shapire, Freund, Bartlett, & Lee, 1998) and bagging (Breiman L., 1996) of classification trees. In boosting, successive trees give extra weight to points incorrectly predicted by earlier predictors. In the end, a weighted vote is taken for prediction. In bagging, successive trees do not depend on earlier trees — each is independently constructed using a bootstrap sample of the data set. In the end, a simple majority vote is taken for prediction.

Ho (Ho, 1995) proposed random forests (RFs), which was later extended by Breiman (Breiman L., 2001). RFs add an additional layer of randomness to bagging. In addition to constructing each tree using a different bootstrap sample of the data, random forests change how the classification or regression trees are constructed. In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node. This somewhat counterintuitive strategy turns out to perform very well compared to many other classifiers, including discriminant analysis, support vector machines and neural networks, and is robust against overfitting (Breiman L., 2001). In addition, it is very user-friendly in the sense that it has only two parameters (the number of variables in the random subset at each node and the number of trees in the forest), and is usually not very sensitive to their values.

The random forests algorithm (for both classification and regression) is as follows:

1. Draw n bootstrap samples from the original data.
2. For each of the bootstrap samples, grow an unpruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample m of the predictors and choose the best split from among those variables. (Bagging can be thought of as the special case of random forests obtained when $m = p$, the number of predictors.)

3. Predict new data by aggregating the predictions of the trees (i.e., majority votes for classification, average for regression).

An estimate of the error rate can be obtained, based on the training data, by the following:

1. At each bootstrap iteration, predict the data not in the bootstrap sample (what Breiman calls "out-of-bag", or OOB, data) using the tree grown with the bootstrap sample.
2. Aggregate the OOB predictions (On the average, each data point would be out-of-bag around 36% of the times, so aggregate these predictions). Calculate the error rate, and call it the OOB estimate of error rate.

The experience has been that the OOB estimate of error rate is quite accurate, given that enough trees have been grown, otherwise the OOB estimate can bias upward.

- ***Gaussian Mixture Models:***

The basis for most speaker identification and verification systems is the GMM, which is used to represent the speakers. More specifically, the distribution of the feature vectors extracted from a person's speech is modeled by a Gaussian mixture density. For a *D*-dimensional feature vector denoted as *x*, the mixture density for speaker *s* is defined as:

$$p(x|\lambda_s) = \sum_{i=1}^{M} p_i^s b_i^s(x),$$

The density is a weighted linear combination of *M* component unimodal Gaussian densities, $b_i^s(x)$, each parameterized by a mean vector, μ_i^s, and covariance matrix, Σ_i^s;

$$b_i^s(x) = \frac{1}{(2\pi)^{D/2}\left|\Sigma_i^s\right|^{1/2}} \times exp\left\{-\frac{1}{2}(x-\mu_m)^t(\Sigma_i^s)^{-1}(x-\mu_m)\right\}$$

The mixture weights, p_i^s, furthermore satisfy the constraint $\sum_{i=1}^{m} p_i^s = 1$. Collectively, the parameters of speaker *s*'s density model are denoted as $\lambda_s = \{p_i^s, \mu_i^s, \Sigma_i^s\}, i = 1, \dots, M$.

Maximum likelihood speaker model parameters are estimated using the iterative Expectation-Maximization (EM) algorithm. Generally 10 iterations are sufficient for parameter convergence.

- ***i-Vector:***

For applications that involve speech or speaker identification, verification or recognition, i-vector features (Dehak et al., 2011) have been state of the art. For the sake of completeness of comparisons, in addition to DBN and LSTM, we will use this method for classification of voice signals. i-vectors convey the speaker characteristic among other information such as transmission channel, acoustic environment or phonetic content of the speech segment. The i-vector extraction could be seen as a probabilistic compression process that reduces the dimensionality of speech-session super-vectors according to a linear-Gaussian model. The speaker- and channel-dependent super-vector $M_{s,h}$ of concatenated Gaussian Mixture Model (GMM) (Reynolds, 2015) means is projected in a low dimensionality space, named Total Variability space,

$M_{s,h} = m + Tw_{s,h}$.

Where m is the mean super-vector of a gender-dependent Universal Background Model (UBM) (Reynolds, 2002), T is called "Total Variability matrix" and $w_{s,h}$ is the resulting i-vector. Probabilistic linear discriminant analysis (PLDA) (Ioffe, 2006) or a simple cosine calculation can then be used to detect the similarities and classify the i-vectors.

2.4 Data collection and feature extraction

The data we use to train and test our model is gathered from 23 human subjects (9 female and 14 male individuals). Subjects read a set of 22 words five times each while their faces were video recorded and EMG signals were acquired from their facial muscles using gold plated surface electrodes that were connected to Grass amplifiers. The words were displayed to the subjects as a

set of slides on a screen in front of them. Each slide contained a single word and was displayed for four seconds. This is a different and much more difficult person identification problem compared to the work presented by Suresh et al. (2011 and 2014), as each word is spoken in a 4-second interval, which in the majority of samples, is not completely filled with the utterance of the word and may start and/or end with periods of silence. We recorded EMG activity and audio and visual signals for a 4.04 second long interval, during which a subject spoke a word (the 0.04 second was due to a delay in displaying the slides and was consistent throughout the experiment). All three modalities of the data and the data collection process are explained in more detail in Section 2.4.1. Section 2.4.2 will include the details of the feature extraction approach applied to the EMG signals.

2.4.1 The data

The video is used for extracting two types of information: 1) a clip of audio signals with a 44.1kHz sampling rate, which is the standard sampling rate in the MP3 protocol (Figure 5), and 2) a sequence of eight images of the mouth area per word (Figure 6) from which we extract a set of coordinates of 20 key points on the lips on each image (Figure 7) using a feature landmark localization method provided by (Pedregosa et al., 2011).

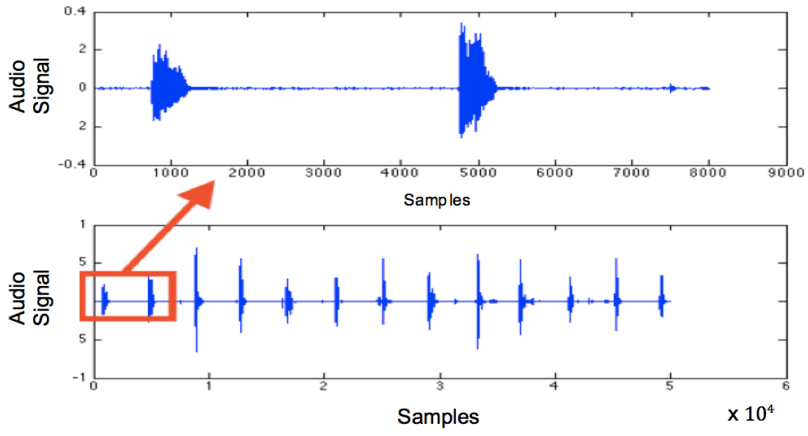

Figure 5. Audio signal for two consecutive words (top) cut from the bottom curve

Figure 6. Sequence of four of the eight images sampled from the video of one of the subjects

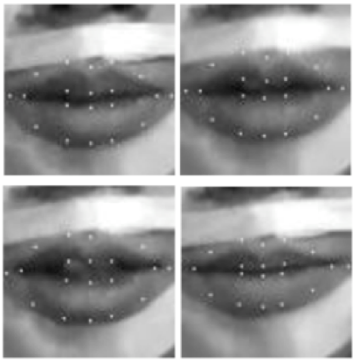

Figure 7. The cropped mouth image with lip key points

The EMG data consists of eight channels captured through eight surface electrodes with a 5kHz sampling rate, then downsampled to 1kHz. Six muscles were chosen including the levator labii superioris, zygomaticus major, risorious, depressor anguli oris, mentalis, and orbicularis oris (Figure 8). These are the major muscles that are involved during speech and have often been used in the literature for speech recognition (Van Boxtel, 2010). We also recorded signals from two other electrodes in order to completely capture all of the movements that occur during speech. One of these electrodes was placed over the mylohyoid muscle, where tongue movements can be measured, and the other one was positioned over the larynx to detect

movements of the vocal chords. The position of the electrodes are illustrated in Figure 8. These electrodes are attached to the subjects' faces as shown in Figure 6. Electrodes placed on the forehead and the left mastoid were used for the ground and reference, respectively. Figure 9 shows EMG signals from all sensors during two spoken words ("buy" and "big").

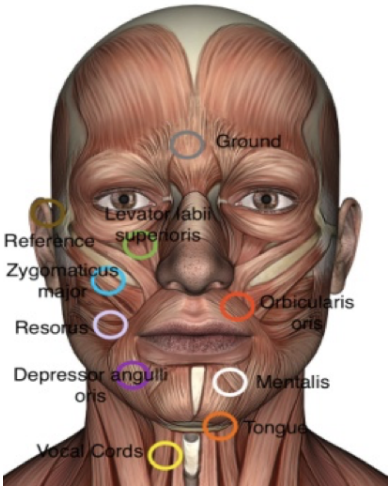

Figure 8. Position of the eight muscles used for capturing the EMG signals, plus ground (forehead) and reference (behind the ear) sensors

We should note here that in many of the current applications on speaker recognition, attaching EMG sensors would be impractical due to the process of attaching the sensors as shown in Figure 6, which could make one question the usefulness of such a system. However, the goal of this study is to explore the potentials of EMG measurements in speaker recognition as a useful biometric and as a complementary source of information or a replacement for other types of sensor data. In other words, we would like to not only see if EMG combined with audio-visual data can improve the performance of speaker recognition compared to audio-visual only, but also to test if EMG can replace the audio and/or visual approach in noisy and/or poorly illuminated environments. The problem of tedious preparation, which is a fair concern in the application of EMG for speaker recognition, can easily be solved to a high extent by using

existing robust portable EMG recording devices, such as the sensors provided by BioSemi, Delsys and SparkFun.

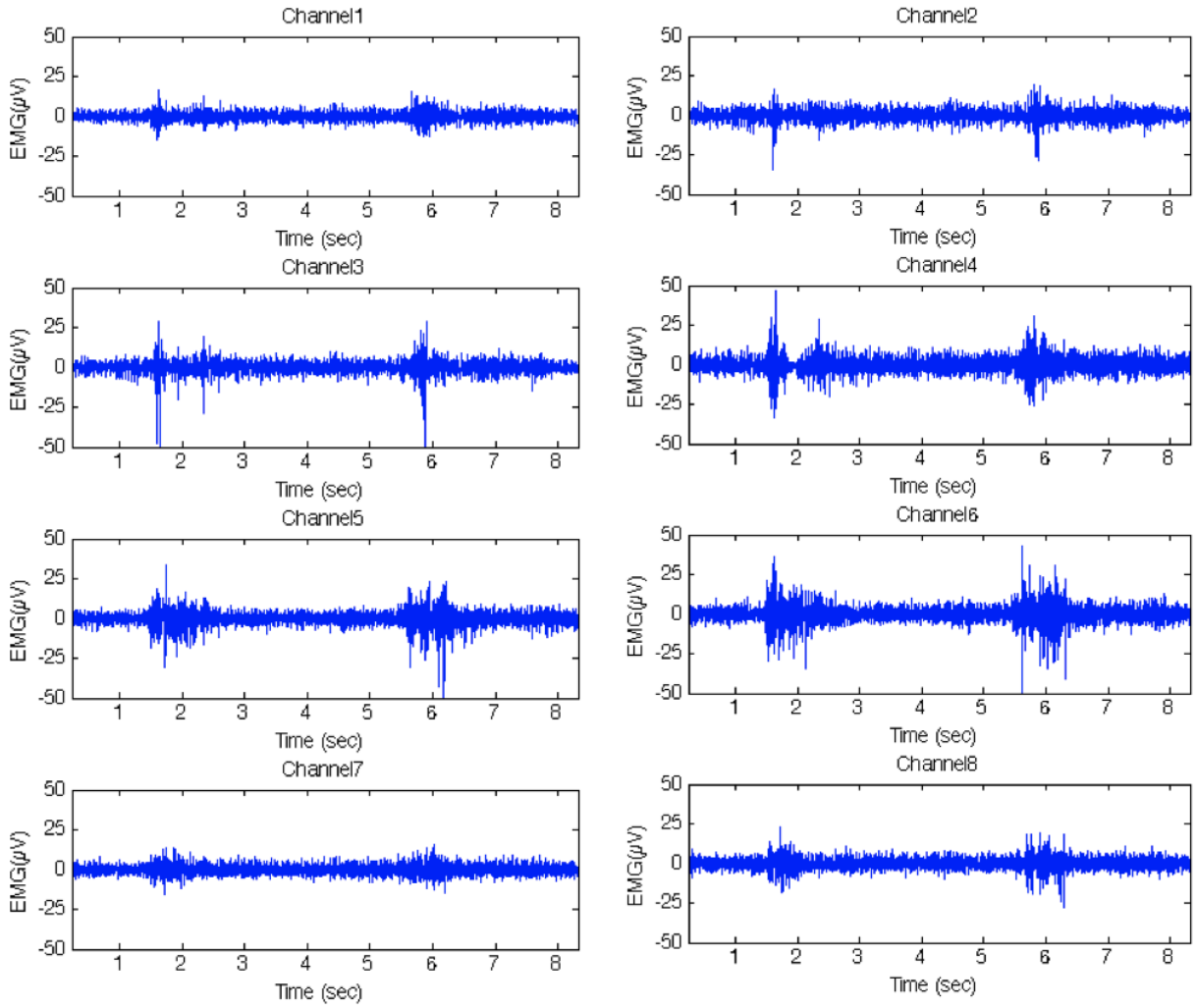

Figure 9. EMG signals from all channels for two consecutive words, buy and big

Another important note is that although the tape used for attaching the EMG sensors is covering parts of the face, it will not interfere with the process of speaking and data collection. The tape is used to secure the electrodes in a way that the human subject is able to move the facial muscles completely and naturally, and the parts of the face that are covered are not used in visual data extraction as only the area around lips is of interest in our work. However, if we want

to use more visual information from the images (for example, the image of the mouth area instead of the coordinates of the points for speaker recognition, or even the whole face images for facial expression recognition), this may cause issues. Using less visible and more adhesive and secure tape would guarantee that this type of visual data could also be extracted and used easily. One solution that we later used in our data collection is to attach the sensors using transparent tape, which makes the face more visible. Also we attached all electrodes on one side of the face to make sure the other side was completely visible.

2.4.2 Feature extraction

Each sample from the EMG dataset is very high dimensional and using it as a raw input modality to train the model is inefficient and impractical. The solution to this problem is to extract features from the EMG signals to reduce their dimensionality as much as possible, without losing valuable information. Many transformation and feature extraction methods have been applied to EMG in different existing works. The most common categories of these methods are summarized in Sharma et al. (2012).

Among all feature extraction techniques applied to EMG, Wavelet Transform (WT) has achieved the best results for classification tasks, especially because it is successful in analysis of non-stationary signals and EMG falls into that category (Khushaba et al., 2011; Phinyomark et al., 2011; Wang et al., 2006). For that reason, we have chosen WT as the feature extraction method in our experiments.

WT methods are categorized into two types: discrete (DWT) and continuous (CWT). We have used DWT due to the discrete nature of the EMG signals. DWT iteratively transforms the signal into multi-resolution subsets of coefficients. Similar to any time-frequency transformation, DWT needs a suitable Wavelet basis Function (WF). Applying WT repetitively to the EMG

signal would normally yield a higher dimensional feature vector than the raw EMG. Thus, selection of an optimal subset of features is an essential step in wavelet analysis. Since WT generates the useful subset of the frequency components or scales of the original signal, picking the most effective subset of extracted features is easy (Phinyomark et al., 2011).

A thorough investigation of different levels of transformation and selection of the WF is performed in Phinyomark et al. (2011). According to the result of those investigations on EMG data, the best WF and decomposition level is found to be the seventh order of Daubechies wavelet (Daubechies, 1992) and the fourth level respectively. We filter the EMG signals to 20-450 Hz bandwidth and used similar settings for the two parameters of WT in our experiments, except that our classification accuracy is maximized when features extracted in level 1 and 4 are used (Figure 10).

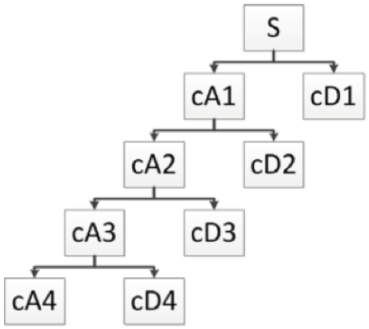

Figure 10. DWT decomposition tree for decomposition level 4 (Phinyomark et al., 2011)

2.5 Analytical experiments

2.5.1 Unimodal DBNs vs. baseline methods

To show the effectiveness of our proposed DBN-based model, we compare its performance on each modality with four baseline methods: GMM, i-vector+PLDA, SVM and RF classifiers. Each method is applied on visual, audio and EMG data.

As mentioned before, the EMG signals are sampled with a 5kHz rate during a 4.04 second interval per word. To reduce the issue of unsynchronized EMG signals due to different points in time at which the speakers start speaking the words, and to obtain more essential characteristics of the signals, we transform the EMG signals into the frequency domain by applying WT to each channel as explained in Section 2.4. Then the wavelet coefficients of the signals from all 8 channels are concatenated to form a single vector of size 528 (=8×66).

For the audio data, we applied MFCC, a popular audio signal feature extraction method that focuses on the frequency band of human audition and is proved to be effective in human speech recognition (Barua et al., 2014, Mohan and Babu N., 2014). We use a Hamming window of size 20ms with a 10ms offset, following the work presented by (Wang, 2014). From each 20ms interval, we then extract MFCC features and choose the 20 most significant coefficients. Then i-vector is used to compute an utterance model (a vector) using the corresponding MFCC features.

The visual features are generated by tracking 20 key points around the mouth on 8 frames sampled from the video during each word. The key points are extracted from these 8 images using the Dlib C++ Library (King, 2009). This results in a list of 160 2D points per word, which are concatenated to form a single vector. The length of each sample vector is 320 (=8×20×2).

With all of these treatments, we have three types of modalities: EMG WT, Audio MFCC and Lip Motion. It is important to note that all these modalities are sequences of samples extracted from signals that are of similar nature: voice and EMG signals have been treated similarly in the literature and the same modeling and classification techniques have been applied to both. GMMs and i-vector-based methods have mostly been used in applications that involve speech or speaker identification, verification or recognition, but have also been utilized to model EMG signals in

the literature (Suresh et al., 2011). For the implementation of GMM and i-vector+PLDA, we used the MSR Identity Toolbox (Sadjadi et al., 2013).

Classifying the modalities can be treated as a pattern matching problem and solved using classification algorithms such as SVM and RF, as proposed in Quitadamo et al. (2017) and Liarokapis et al. (2013). The key points captured from the mouth area form a sequence that can be modeled using similar family of distributions and classified by the same methods.

All unimodal DBNs consist of four layers. The number of units in the input layer is set equal to the length of the training samples. Different number of units taking values from the set {50, 100, 150, 200, 250} were tested for the first and second hidden layers throughout the analyses and the combination that maximized the majority of the results was 250 and 100 units for the first and second hidden layer respectively. In order to always keep the parameters of the DBN models consistent for a fair comparison of the results, we use the same combination (250 and 100) across all analyses. The output of the DBNs is represented using the one-hot encoding technique for the class labels. Thus, the number of units in the output layer is equal to the number of classes (speakers), which is 23.

We are dealing with a small dataset in our analyses. This is a common situation with biometrics data involving EMG or other similar modalities such as electroencephalography (EEG) due to their expensive and time-consuming collection process. To make up for this issue, we repeat each of the analyses 100 times. The confusion matrices and accuracies are then averaged over all 100 iterations and paired t-tests are performed to compare the methods and assess the statistical significance of the differences observed in their accuracies. We have used ttest in MATLAB for this test which follows the format ttest(x,y) and returns a test decision, h, for the null hypothesis that the data in x – y comes from a normal distribution with mean equal to

zero, using the paired-sample t-test. The result h is 1 if the test rejects the null hypothesis at the 5% significance level, and 0 otherwise. In our case, x and y are each a vector of accuracies of a classifier from 100 iterations. We report the result of the t-test as t(*df*) = *tstat*, p = *p-value*, where *df* is the degrees of freedom of the test, *tstat* is the value of the test statistic, and *p-value* is the probability of observing a test statistic as extreme as, or more extreme than, the observed value under the null hypothesis.

For all methods, in each iteration, we divide the data into two sets: the training set, which contained 60% of the samples, or three out of five repetitions of each word spoken by each subject, selected randomly. The remaining 40%, which contain two samples of each word spoken by each subject, was used as the test set. This holds for all modalities. In each analysis, once the three training samples are chosen randomly from the set of five repetitions of a word, the same training and test sets are used across all methods for a fair comparison. Figure 11 shows the accuracy of GMM, i-vector+PLDA, SVM RF, and unimodal DBNS applied on each modality, with both their mean accuracy values (%) and their standard deviations. As seen in Figure 11, DBN outperforms the majority of classifiers, with the smallest standard deviations.

SVM settings did not work well for either audio MFCC or lip motion, whereas RF did not work well on EMG WT. For Audio MFCC, we realized by the results of our experiments that a SVM with kernel would be too complicated for this data and reduces the accuracy, so we tried a linear SVM, which still does not perform well (31.8%). In general, SVM is not suitable for problems with 1) too many classes, 2) huge number of features (size of each sample) compared to the number of samples, and if 3) the model needs to be trained on the entire dataset in parallel and we need to account for the knowledge provided by all samples simultaneously.

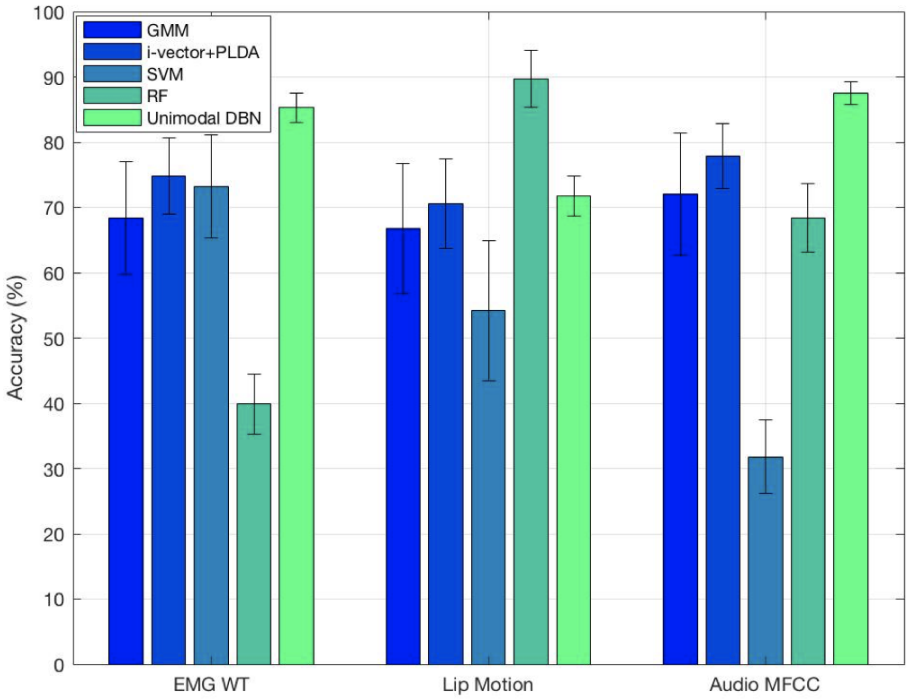

Figure 11. The mean and standard deviation of the accuracies of the baseline approaches compared with the unimodal DBN for speaker recognition over 100 iterations

Voice is the superior modality for speaker recognition when available, and for that reason, it has been used as the primary source of information in the majority of speech and speaker recognition research. However, using only voice might not be effective to separate two speakers, especially when only a single word is spoken. The confusion matrices in Figure 12 show that in areas where voice is not powerful and informative enough to correctly classify the samples, EMG can compensate for it, and vice versa. As an example, speaker #4 is classified with 15.9% error rate using voice, but the error rate using EMG is 8.1% for this speaker. An opposite example is speaker #12, which is classified with an error rate of 13.7% and 6.2% using EMG and voice, respectively.

This leads to the conclusion that EMG and voice, or theoretically any two modalities that capture different types of information about the data, are complementary and can make up for shortcomings of one another. Given the high overall accuracies of EMG and voice-based DBN,

EMG proves to also be a powerful modality for speaker recognition. In addition, these two modalities not only can complement each other (which is investigated in the next sections), they can replace each other as an alternative source of information if need be, e.g. when one is unavailable, very noisy or corrupt. The characteristics captured by different modalities are useful both for discriminating the speakers, as well as detecting similarities between them.

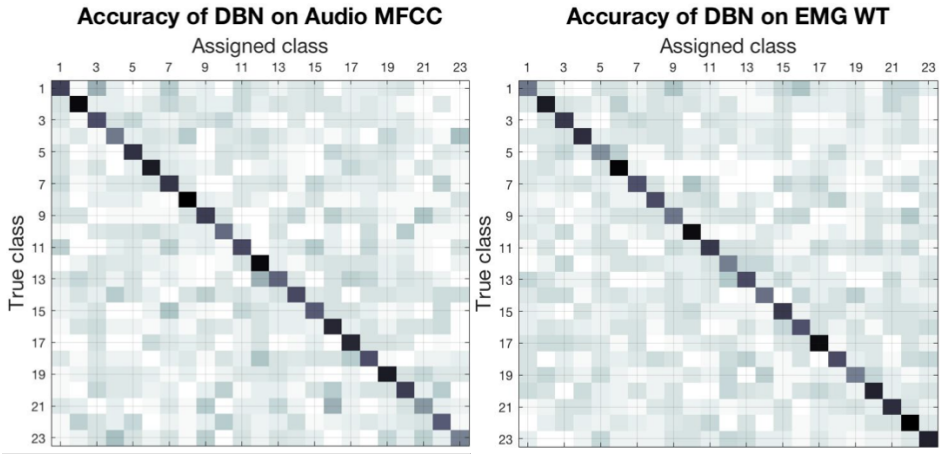

Figure 12. Confusion matrices of unimodal voice (left) and EMG-based (right) DBN classifiers. Color levels are adjusted for illustration purpose

As mentioned earlier, DBN outperforms the majority of classifiers, except in one case where RF has a higher accuracy. To confirm the significance of the improvement obtained by DBN in cases where DBN outperforms other methods, we compared the DBN with the next most accurate classifier using a t-test for each modality. DBN performed significantly better than i-vecor+PLDA when classifying the EMG WT data, $t(99)=3.173$, $p=7.4\text{x}10^{-4}$. For Audio MFCC, DBN was significantly better than i-vecor+PLDA, $t(99)=3.461$, $p=3.2\text{x}10^{-4}$. These statistically reliable differences demonstrate the power of DBNs over the other methods.

2.5.2 Multimodal deep belief networks and experimental results

The results of Section 2.5.1 lead us to three conclusions: 1) Overall, DBNs are stronger models for classifying our dataset. 2) Each of the modalities contains valuable non-overlapping

information about the data and thus, the next interesting analysis to try is classifying the speakers based on the combination of modalities using a multimodal DBN structure. 3) DBNs are very stable compared to other classifiers, with lower standard deviations across 100 iterations.

The multimodal DBN consists of two or more unimodal DBNs, where the output (classification) layer of each unimodal model is removed and the second hidden layers of all unimodal models become the input to a single shared layer. The size of this shared layer is equal to the total number of units in all unimodal DBNs. For instance, if we combine two modalities, with the settings explained for the unimodal DBNs, the number of units in the shared layer is 100+100, or 200. Then an output layer is added to the top of the shared layer for classification with 23 units for the 23 classes (i.e., speakers).

We performed three sets of analyses. The first analysis compares various combinations of modalities. The second set verifies if the DBN models are word-independent, i.e. trained and tested on non-overlapping subsets of words. Then, in the third set of analyses, we further test if the DBN models allow mismatches between two modalities as long as they belong to the same person. That is, when the EMG and audio parts come from two different words, spoken by the same person.

2.5.2.1 Multimodal speaker recognition using combinations of modalities

In this part of the analyses, the training and the testing samples are the same as in the baseline approaches for fair comparison. The results achieved by various multimodal DBNs are depicted in Figure 13.

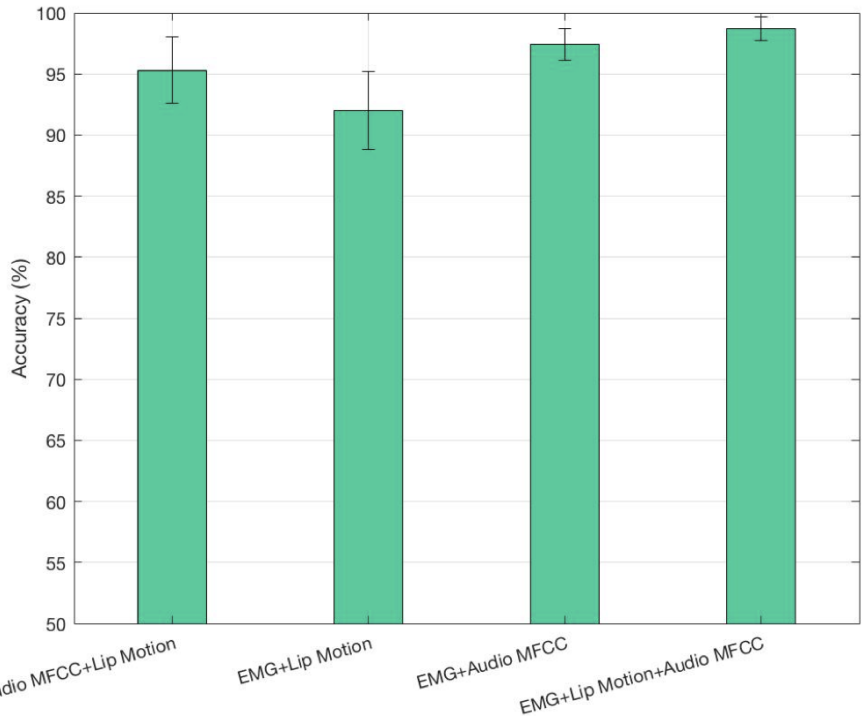

Figure 13. Mean and standard deviation of accuracies of multimodal DBN classifiers for speaker recognition, over 100 iterations

As we expected, combining modalities increases the accuracies of classifiers. In particular, adding EMG significantly boosts the performance of speaker recognition using audio and visual based unimodal DBNs, by roughly 9% and 19%, respectively. This confirms the power of the EMG modality.

The multimodal DBN with all three modalities achieves the highest accuracy, but the contribution of video is only marginal (1.3% improvement). Therefore EMG seems to be a much better modality than lip motion. Intuitively, although EMG and lip motion are captured differently, EMG covers the information provided by lip motion to some extent, as lip motion is the visible result of muscle movements that EMG sensors capture, but only around the mouth area.

Since the accuracies obtained in Figure 13 differ in relatively small amounts, we applied t-tests to ensure the differences are statistically significant and meaningful. The pair of Audio

MFCC + Lip Motion was significantly worse at identifying speakers as compared to the Audio MFCC+EMG, $t(99)=2.903$, $p=7.02\text{x}10^{-4}$, which indicates that EMG is superior to lip motion. For Lip Motion + Audio MFCC vs. EMG + Lip Motion + Audio MFCC, adding EMG to the multimodal DBN significantly improves its accuracy, $t(99)=3.050$, $p=3.31\text{x}10^{-4}$.

2.5.2.2 Word-independent uni- and multi-modal speaker recognition

The previous speaker identification experiment is word-dependent as the training set includes instances of all the words. To see how well the multimodal model performs in a word independent setting, we change the training and test set as follows: Instead of using three samples of a word in the training set and the remaining two samples of the same word in the test set, we choose the first 13 words of the total 22 words for training and the rest for testing. This way, there is no overlap of words between training and test data and the proportion of training to test samples remain very close to 60-40%, similar to previous analyses. To be consistent with previous sections, this experiment is repeated 100 times with randomly selected 13-word subsets as the training set and the remaining 9 words as the test set. We use the same structure for the model as in previous experiments.

Figure 14 shows the results of the analysis. The observations we made in Section 2.5.1 still hold in terms of modalities, and comparing Figure 13 and Figure 14, the performances are highly comparable, indicating that the trained DBN models are word-independent and the DBN models, either unimodal or multimodal, successfully capture the identity of the speakers, independent of the word.

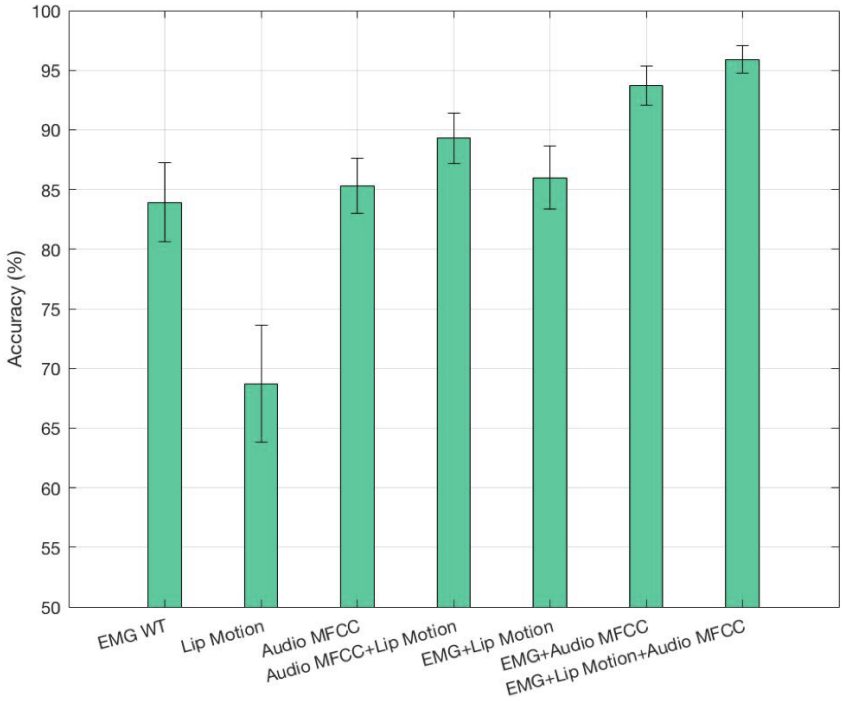

Figure 14. Mean and standard deviation of accuracies of word-independent multimodal DBNs over 100 iterations

The statistical significance of the improvements resulted from combining modalities was assessed using t-test. Specifically, we tested whether adding EMG improves the accuracy of the uni- and bimodal DBNs, i.e. whether unimodal Audio MFCC was statistically any different from bimodal EMG+Audio MFCC. This t-test resulted in a significant difference between these two cases, demonstrating that adding EMG increases speaker identification accuracy, $t(99)=2.691$, $p= 2.88 \times 10^{-4}$.

We also compared the multimodal DBN (EMG WT + Lip Motion + Audio MFCC) and the most accurate bimodal DBN (EMG WT+ Audio MFCC). This difference was also statistically reliable, yields $t(99)=2.331$, $p=1.43 \times 10^{-3}$, indicating that although Lip Motion has the lowest accuracy among other unimodal cases, combining it with other modalities further improves the overall accuracy of speaker recognition.

2.5.2.3 Unsynchronized multimodal speaker recognition

In this section, encouraged by the results of the analysis performed in previous section, which suggest that speaker identification is not dependent on the contents of EMG or audio, we verify our assumption that the EMG and audio (or lip motion) parts of the data do not need to match. This will make the multimodal data collection more flexible in that the audio (or visual) and the EMG data do not need to be collected in a synchronized way. For this purpose, we tested if the models would still work well if we choose the EMG and audio parts from different samples of the same person in our dataset. We randomly paired the EMG with the audio, lip or lip+audio data of the same person and we generated the training and test sets by selecting 60% and the remaining 40% samples respectively. The results of this analysis are shown in Figure 15.

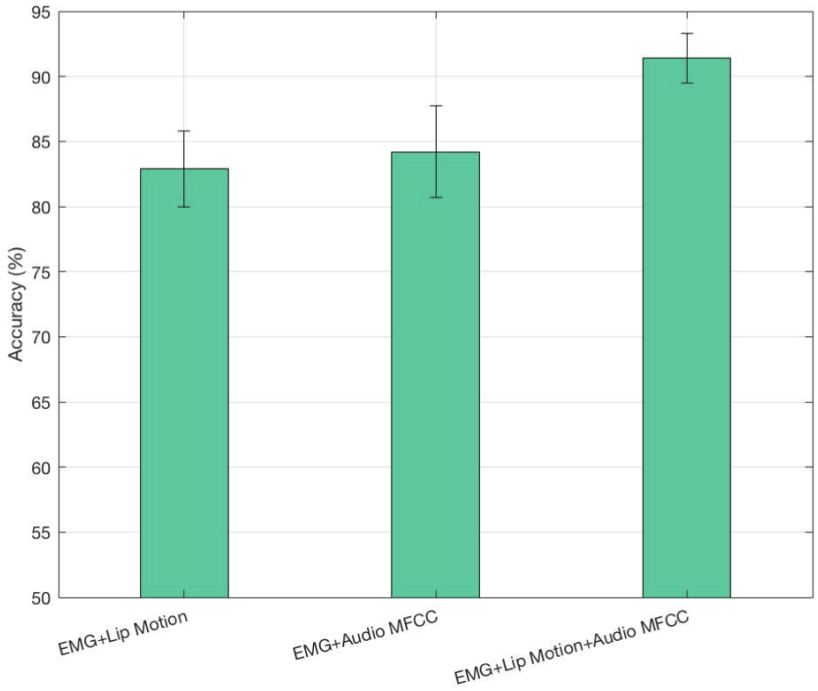

Figure 15. Mean and standard deviation of accuracies of speaker recognition with asynchronized and unmatched multimodal data, over 100 trials

Comparing the results in Figure 15 with Figure 13 and Figure 14 indicates that performance is still very strong. Intuitively, the multimodal DBN attempts to learn the association between the

modalities. This association included the “dependence” of these modalities in previous analyses. In the current analysis on the other hand, the association is the “independence” of the modalities, given the speaker. The results indicate that although the accuracy of classification is lower in this case, it is still very strong and comparable with previous analyses. This further implies that the trained DBN models for speaker recognition are both word-independent and content-independent.

We note that the audio is the most content-dependent modality among all three modalities used in our work. Two different words may look very similar when only visual information is taken into account, despite their different corresponding audios. This explains why the combination of Audio MFCC with EMG shows only 1.3% improvement over the combination of EMG and Lip Motion.

The word independent and random matching multimodal settings are practically more useful. As opposed to a word dependent setting where the person has to speak the same word during training and test of the system, the random matching experiment is a more versatile setting, as there is no need for the EMG and audio/lip motion data to come from the same word, nor captured at the same time.

2.6. Conclusions

To recognize the speaker based on EMG, audio, and visual data, we propose a DBN-based multimodal approach. The role of a unique sensor modality, namely the EMG data, is demonstrated for speaker recognition. Compared with baseline approaches, i.e. GMM, i-vector+PLDA, SVM and RF classifiers, the DBN model is more robust to all types of features, except when RF was applied to lip motion, and the multimodal approach can outperform all

individual features. The EMG modality significantly boosts the speaker recognition rate and can replace or complement the audio modality if the latter is not available or is extremely noisy. We have further shown that the DBN models are word-independent and different parts of the multimodal data do not need to match, as long as they belong to the same speaker.

As we have noted, with the current technology of EMG sensors, using EMG in real speaker identification/verification applications might not be very practical. However, the use of EMG could have great potential in some critical applications such as forensic analysis. Furthermore, this study may also raise interest in the design of more user-friendly EMG sensors. With the advances in sensor technologies, more portable EMG sensors might be ready in the near future. However, even the most portable EMG sensors are not practical in everyday applications due to the fact that the sensors have to directly touch the skin to capture the signals. A future direction could be to explore the effectiveness of incorporating EMG data only in the training of more robust and descriptive models for speaker recognition and other recognition tasks using audio only or audio plus visual data.

Chapter 3

Analysis of Different Modalities for Emotion Recognition

3.1 Introduction

Emotional state can be observed or measured in many different ways, including through facial expressions, speech, and physiological signals. The idea of emotion recognition while speaking has been investigated by several researchers in applications such as human-computer interaction (HCI) and call center monitoring. These applications have also produced multiple datasets that are being used by researchers. The goal of the majority of emotion detection work has been to optimize the accuracy of emotion recognition, more recently by utilizing the state-of-the-art statistical or machine learning models and the most relevant modalities such as visual information, vocal features, body movements and posture, or physiological signals. Several attempts have been made to combine multiple modalities to further improve the accuracy of the emotion recognition models.

The goal of the current research was twofold. First, we examined the efficacy of using different modalities and machine learning models for emotion recognition. We collected a rich new dataset by recording video, audio, facial muscle movements (with EMG signals), and brain activity (with EEG signals) while subjects (actors) spoke a generic sentence expressing one of the seven different emotions. We then applied a set of state-of-the-art feature extractors, each

suitable for a specific modality, before applying the most effective deep learning and statistical models. Various different machine learning models were compared to determine the best models for the different modalities.

The second goal, which is the main focus of this work, was to compare the characteristics that are specific to each modality and the conditions in which each modality performs optimally . We analytically show that even though each modality might seem ineffective in some settings, if used correctly, their unique contributions to emotion recognition can be effectively applied to increase classification. Scientifically, we investigated, using the same dataset, how much spatial-temporal visual facial expression, auditory speech information, facial muscle movements, and neural activity can classify a person's emotion and in what stages of speech and expression that each modality best captures the emotion. To assess how neural activity may be used to detect emotions, we used EEG to record brain activity while subjects expressed different emotions. Thus, the overall goal of this work was not to improve existing emotion recognition methods, but to thoroughly study different emotional state detecting modalities and to determine the optimal stages of information in the signals, the best categories of feature extractors, and the most appropriate machine learning models that would best fit each modality.

As a summary, the contributions of the work include: (1) A new multimodal dataset was collected with four different modalities: audio, video, EMG, and EEG. (2) A thorough comparison was performed across these four modalities to determine how to optimize the data, which features and models are most informative, and to offer insights into the effectiveness of each modality to classify emotions. (3) Most notably, we study at what stage of each modality, especially for the neural activity measured with EEG signals, emotion information prevails and for how long they remain reliable.

3.2 Related work

There is a large body of work on emotion recognition using different modalities separately and in combination with one another. Three categories of datasets are usually used in analyzing emotions: acted emotions, natural spontaneous emotions, and elicited emotions (Kessous et al., 2010). Although different actors may understand and interpret instructions differently and may actually experience the emotions to different degrees, data obtained from acted emotions are less ambiguous because actors express the exact emotions they were instructed to act.

In contrast, spontaneous speech and emotions can, for example, be collected from call center data (Gupta & Rajput, 2007), or through human-computer interaction (Fragopanagos & Taylor, 2005). These emotions are more diversified and are often difficult to classify given that the data must be mapped onto a limited number of classes. Even if it is evident that emotion research should ideally target natural databases, acted databases are more systematically controlled and useful, especially neural activity will be measured. Furthermore, there is a direct correspondence between the collected data and their labels, generally resulting in higher accuracy in emotion recognition (Vogt & André, 2005; Burkhardt et al., 2005). We therefore uses acted emotions for data collection in this work.

Generally, facial expressions and speech have been the two most used modalities for emotion recognition, although other modalities have also been investigated. In the area of unimodal emotion recognition, there have been many studies using a variety of different, but single, modalities. More recently, several attempts to emotion recognition from multimodal data were made. Some examples of multimodal emotion databases can be found in (Kessou et al., 2010; Soleymani et al., 2012; Zhang et al., 2016). There have been many studies using different single modalities. Facial expressions (Shan et al., 2009; Li et al., 2017; Mollahosseini et al., 2017),

vocal features (Yacoub et al., 2003; Vogt et al., 2008; Parlak & Diri, 2013), body movements and postures (Gunes & Piccardi, 2007; Crane & Gross, 2007; Bernhardt, 2010; Piana et al., 2014), physiological signals such as skin temperature, skin conductance, blood volume pulse and heart rate (Gouizi et al., 2011; Uma, 2014) and EMG (facial muscle activity) (Lahane & Sangaiah, 2015; Spampinato et al., 2016; Palazzo et al., 2017) have been used as inputs during these attempts. Nevertheless, most of the work has considered the integration of information from facial expressions, speech and body gestures, as many psychological studies have highlighted the need to consider the integration of multiple modalities for a proper inference of emotions (Bänziger et al., 2009; Kessous et al., 2010; Piana et al., 2014; Liu et al., 2016).

Another approach that has more recently been explored for emotion recognition is through EEG, which measures electrical activity in the brain (Kothe et al., 2013; Blaiech et al., 2013; Lahane & Sangaiah, 2015; Spampinato et al., 2016; Palazzo et al., 2017). EEG is especially interesting due to its capability to detect internal emotional states, as opposed to the other modalities mentioned above. Some previous studies (Kothe et al., 2013; Blaiech et al., 2013; Lahane & Sangaiah, 2015; Spampinato et al., 2016; Palazzo et al., 2017) have incorporated the use of EEG in attempts to determine the inner emotional (affective) state. Here, we recorded EEG signals during different expressed emotional states and compared them with other modalities.

3.3 Deep learning approaches and baseline methods

Different machine learning techniques have been used in emotion recognition. One approach, which has been successful, is to use deep learning approaches because they have the ability to learn the most relevant features with respect to the task. Three deep learning models are used in

this chapter: Convolutional Neural Network (CNN) (LeCun, 1995), Long-Short Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) and Deep Belief Network (DBN). Although all three models are characterized as being 'deep', as they use layers of latent or hidden variables, they have very different characteristics.

The architecture of deep learning techniques can be categorized into two different categories: generative and discriminative. The deep models that fall into each category often share the properties of the other category, making it difficult to draw a clear boundary between the two groups of models. Generative models are very useful for both classification and regression tasks, especially when data preparation and pre-training of the parameters of the model are necessary. These models have the ability to initialize the search through the parameter space in an area that potentially contains the solution. On the other hand, the architecture of the discriminative models has direct ability to classify the data (Giri et al., 2016). In other words, the former models describe the distribution of data, whereas the latter models describe the distribution of targets conditioned on data (Deng & Jaitly, 2015). In the current chapter, we investigated how the two different models handle the data from the four different modalities and draw some useful conclusions about the effectiveness of these models for different types of datasets.

Example of discriminative architectures include CNN, Recurrent Neural Network (RNN) (Britz, 2015), and LSTM. The DBN, which has been explained and used in the previous chapter, is an example of generative models. The DBN has been explained in detail in section 2.3.1. We will now review the CNN and LSTM models.

The idea of LSTM goes back to the RNN model and its shortcomings. RNN is an artificial neural network model that has feedback connections in the hidden units. Because the previous

states in the hidden units are used as inputs, RNN can store historical information like memory and can solve context- dependent tasks with the architecture. The gradient based training of RNN has a problem that derivatives propagated via recurrent connections become too small or too large, due to multiplicative gates in the units. This vanishing and explosion gradient problem makes learning of RNN difficult.

LSTM is a special type of RNN architecture to overcome the vanishing gradient problem. The schematic view of LSTM is shown in Figure 16. LSTM units receive external inputs and generate hidden outputs. LSTM consists of three gates (input, output, and forget gates) and a memory cell. The gates and memory cell are internally connected with weighted links, and the gates are also connected with external sources, which are current sequential inputs, x_t and previous hidden states, h_{t-1}. The hidden output, h_t , is calculated from x_t , h_{t-1} , and previous state of the memory cell, c_{t-1}.

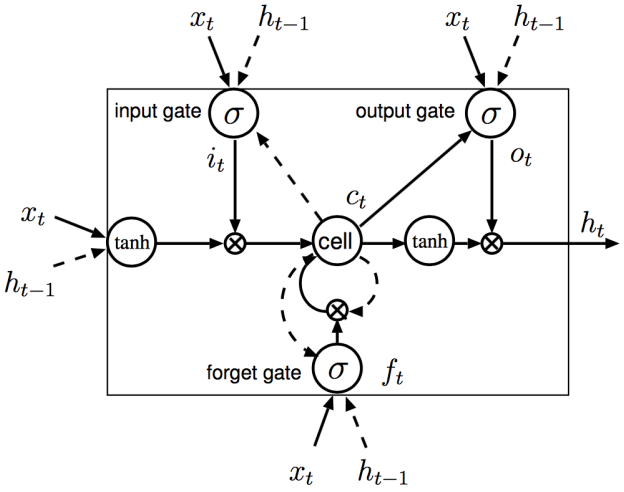

Figure 16. LSTM unit (Noguchi et al., 2016)

The equations of LSTM can be expressed as follows.

$$
\begin{aligned}
\boldsymbol{i}_t &= \sigma\left(\mathbf{W}_{xi}\boldsymbol{x}_t + \mathbf{W}_{hi}\boldsymbol{h}_{t-1} + \mathbf{W}_{ci}\boldsymbol{c}_{t-1} + \boldsymbol{b}_i\right) \\
\boldsymbol{f}_t &= \sigma\left(\mathbf{W}_{xf}\boldsymbol{x}_t + \mathbf{W}_{hf}\boldsymbol{h}_{t-1} + \mathbf{W}_{cf}\boldsymbol{c}_{t-1} + \boldsymbol{b}_f\right) \\
\boldsymbol{c}_t &= \boldsymbol{f}_t\boldsymbol{c}_{t-1} + \boldsymbol{i}_t \tanh\left(\mathbf{W}_{xc}\boldsymbol{x}_t + \mathbf{W}_{hc}\boldsymbol{h}_{t-1} + \boldsymbol{b}_c\right) \\
\boldsymbol{o}_t &= \sigma\left(\mathbf{W}_{xo}\boldsymbol{x}_t + \mathbf{W}_{ho}\boldsymbol{h}_{t-1} + \mathbf{W}_{co}\boldsymbol{c}_t + \boldsymbol{b}_o\right) \\
\boldsymbol{h}_t &= \boldsymbol{o}_t \tanh\left(\boldsymbol{c}_t\right)
\end{aligned}
$$

where σ is the sigmoid function, *i*, *f*, *o*, *c* and *h* are the input, forget, output gates, memory cell and hidden activation vectors, respectively (Graves, 2012; Noguchi et al., 2016)

In addition to the deep learning methods explained in this section, we will use i-vector as the baseline method for classifying the voice modality. This method has been explained in section 2.3.3.

3.4 Data Collection and Feature Extraction

The main steps of data collection for this part of our work is very similar to the previous chapter, except that the goal of this chapter is to explore the effect of different modalities in emotion recognition, with more emphasis on the unique characteristics of the EEG with respect to the emotional state. In addition, we have tried different feature extraction methods and deep learning models in this chapter.

3.4.1 The data

The data we used to train and test the models was gathered from 12 human subjects (5 female and 7 male individuals), who participated after informed consent. The study was approved by the Institutional Review Board of the City University of New York. In general, having a relatively small number of subjects is typical in neuroscience studies due to the difficulty in data collection. For this study, each testing session lasted approximately two hours and it took approximately four months to recruit the actor subjects and collect all of the data. We included a large number

of repetitions/instances in the dataset to minimize variability. For each of the 7 emotions, 50 trials were expressed for a total of 350 emotion instances per subject. We will release the dataset following publication of the paper corresponding to this chapter for research purposes. The subjects either had acting experience or were acting students because the emotions needed to be expressed as naturally and believably as possible. Every five seconds, one of the seven standard emotion labels was presented on a monitor placed 57 cm in front of the subject. The emotions were happiness, sadness, anger, surprise, fear, disgust, and neutral. Each time an emotion label appeared on the screen, the subject uttered the sentence "The sky is green" while trying to mimic the facial expression and experience the emotion associated with that label. This sentence was chosen because of its neutral content, thereby minimizing interference with any emotion that the subject was trying to experience and express. During the utterance, the subject's face and voice were video recorded and EMG and EEG signals were acquired from their facial muscles and scalp using gold plated surface electrodes that were connected to Grass amplifiers. The camera and microphone were placed in front of the subject to ensure an adequate quality of the acquired video and voice.

Each emotion label was displayed for 4 seconds, and a one-second break was given between every emotion. Overall, the longest it took the subjects to speak the sentence was approximately 2.5 seconds. The entire interval, therefore, was not completely filled with the utterance of the sentence and started and/or ended with periods of silence.

The entire session was divided into five sub-sessions. Each sub-session contained 10 repetitions of each emotion in random order. That is, we used a 7x10 total number of emotions per sub-session, or 5x7x10=350 emotions overall, for each of the 12 subjects. Between every two consecutive sub-sessions, the subject took an optional break that was arbitrarily long.

All four modalities used in our analyses (images, voice, EMG and EEG) and the data collection process and the details of the feature extraction approaches applied to the modalities are explained in Section 3.4.2.

3.4.2 Feature extraction

Similar to the previous work on speaker recognition, the video was used for extracting two types of information: 1) a clip of audio signals with a 44.1kHz sampling rate, and 2) an image sequence of 24 screenshots per sentence. The 24 images were evenly sampled from the 2.5 seconds (on average) during speech, such that this window included most of the emotional expression. Only 24 frames were used for computational efficiency, following the work presented in (Li et al., 2017). A few samples of the screenshots are shown in Figure 17.

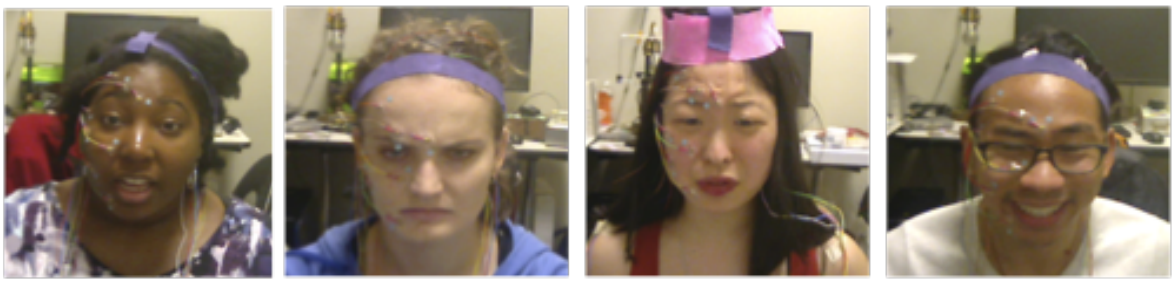

Figure 17. Screenshots from four videos

The audio was recorded using a laptop's microphone with the 44.1kHz sampling rate. We divided the audio into 20ms intervals with 10ms offsets and then extracted MFCC features from each interval separately. The features extracted from the intervals formed a sequence that embed both frequency and time information.

The EMG data consisted of six channels captured through six surface electrodes with a 5kHz sampling rate that was then downsampled to 1 kHz. Six muscles were chosen: the depressor

anguli oris, zygomaticus major, levator labii superior alaeque nasi, levator labii superioris, procerus, and occipitofrontalis (Figure 18). These are the major muscles that are involved during facial expressions and their equivalent facial Action Units (AUs) have often been used in the literature for facial expression recognition (Li et al., 2017; Wang et al., 2013; Katsikitis, 2012; Zeng et al., 2009). A band-pass Butterworth filter (20 to 450 Hz) was applied to the EMG data to eliminate noise and meaningless parts of the signals.

EEG data were acquired using the same sampling frequency as the EMG data but through 8 surface electrodes placed onto the scalp: F3, Fz, F4, Cz, P3, Pz, P4, and O2 (Figure 19). The preprocessing steps applied to the EEG data were similar to the EMG data but with different bandpass filter settings (0.1 to 30 Hz). Figure 20 shows samples of EMG and EEG signals collected from one of the subjects while acting “fear”.

All electrode impedances were below 10 kΩ at the start of the experiment. After filtering the EEG and EMG signals, wavelet transforms (WT) were applied for feature extraction, similar to the previous chapter.

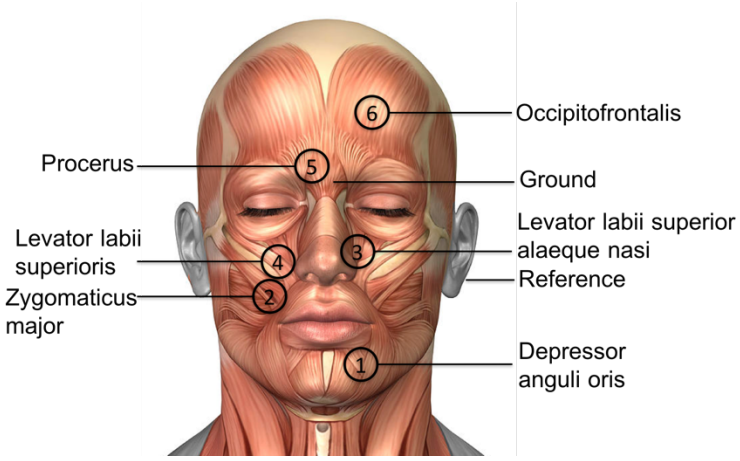

Figure 18. the position of the sensors on the face

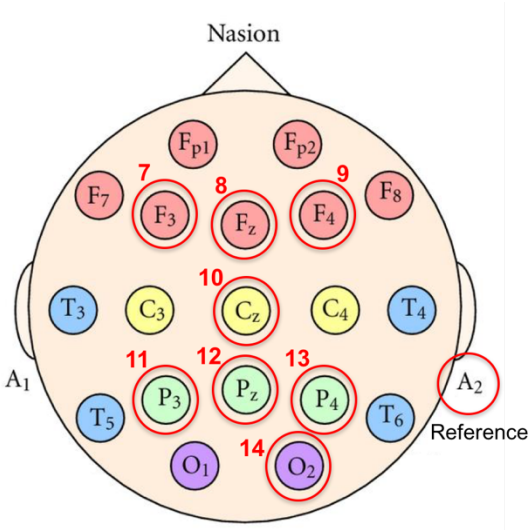

Figure 19. the position of the sensors on the scalp

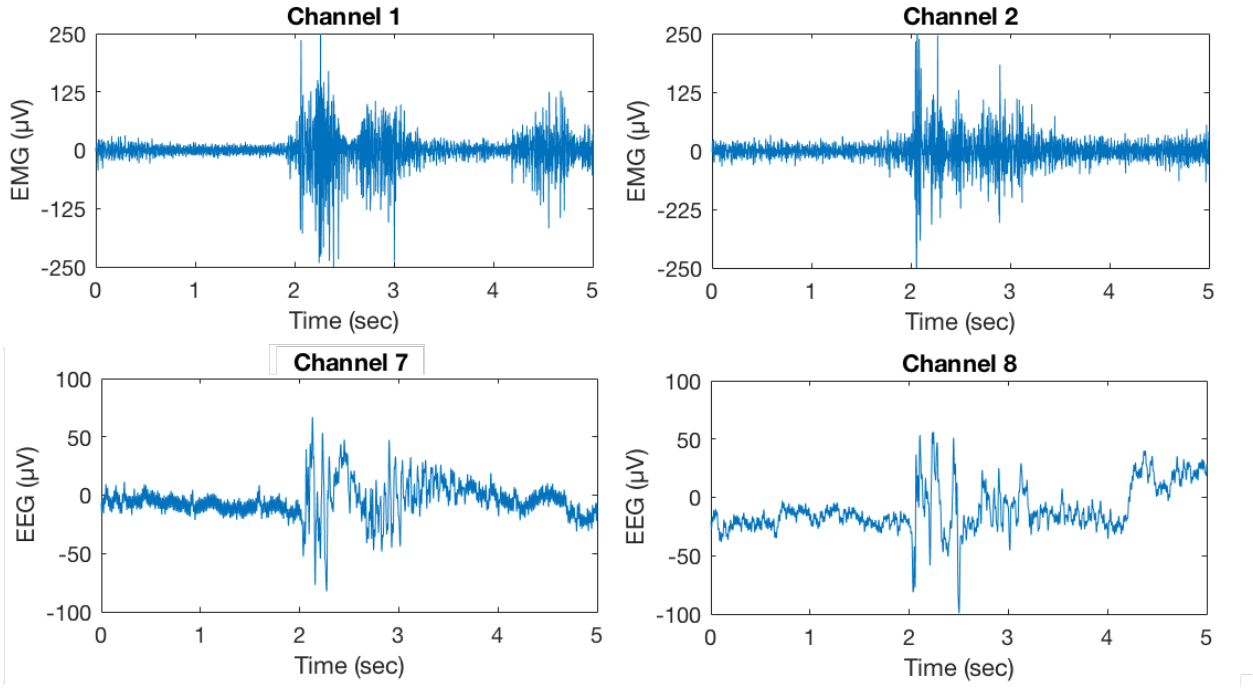

Figure 20. EMG/EEG Data (Subject 03, Emotion Fear, EMG channels 1 and 2, EEG channels 7 and 8)

The features extracted from the images differ from our previous work. To extract emotions from the images, we first cropped the face area on each frame using the open source DLib C++ Library (King, 2009). A few cropped frames from a "happy" sequence are shown in Figure 21.

Each cropped image was 186x186 pixels, which was enlarged to 224x224 pixels to fit into the feature extraction method. Note that transparent tape was used on the faces to minimize occlusion of the facial features. This modification has extremely improved the quality of extracted features from the images, since we are using the entire image for feature extraction as opposed to our previous work where we extracted points around the mouth area.

Figure 21. Face cropping from the video using DLib

Using pre-trained models for visual feature extraction has become a very common and promising approach among researchers (Li et al., 2017; Marmanis et al., 2016; Fang et al., 2015; He et al., 2014; Allen et al., 2006). As explained earlier, we have a sequence of 24 images per utterance of the sentence. The features were extracted by applying an integrated deep learning model with the pre-trained VGG-16 network, followed by the ROI network, as proposed in (Li et al., 2017). The VGG-16 net is developed by the Visual Geometry Group at Oxford (Simonyan & Zisserman, 2014). The VGG team won the first and the second places in ImageNet ILSVRC-2014 competition in the localisation and classification tasks respectively. Since then, their proposed model has been state of the art and is widely used for feature extraction (Li et al., 2017; Fang et al., 2015). The layer configuration of VGG-16 is shown in Figure 22.

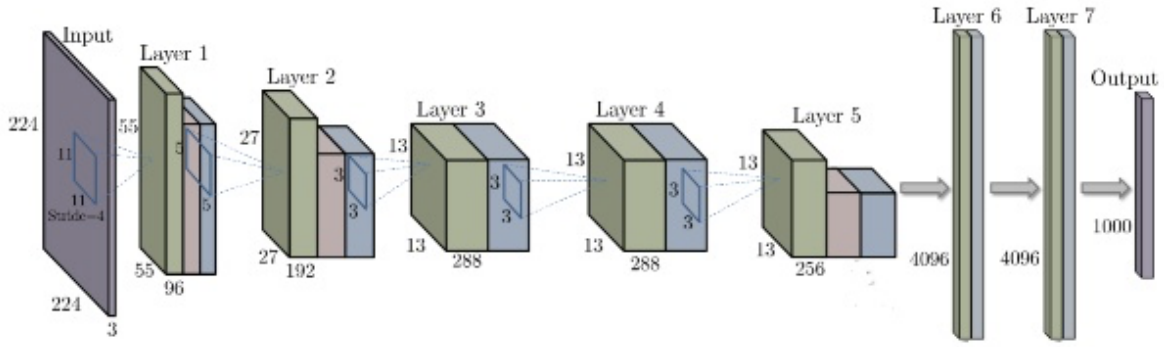

Figure 22. The layers and structure of the VGG-16 net

The reason we chose the VGG+ROI model rather than more sophisticated models, such as ResNet (He et al., 2016), is that the VGG model is sophisticated enough for our data and VGG+ROI has been a further trained model using more than 10K facial expressions. More specifically, the ROI nets are designed to ensure that regions of interest on the faces were learned independently; each sub-region (out of 20 in this case) has a local CNN - an ROI net, whose convolutional filters were only trained for the corresponding region. The structure of the VGG+ROI model is illustrated in Figure 23. The VGG net's output from fully connected layer 7 (fc7) provided the input to the ROI net. Each feature vector was obtained from the output of the last layer and had 2048 elements.

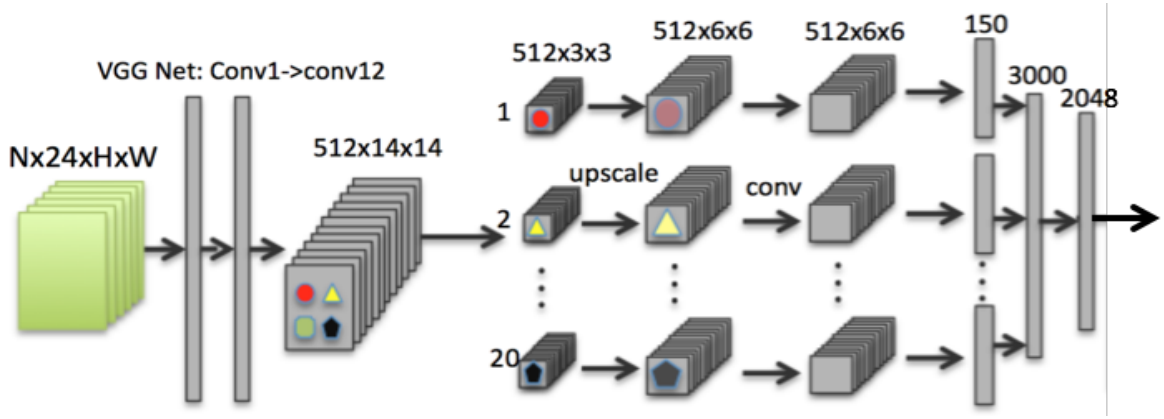

Figure 23. Framework of the proposed neural network with VGG Net and ROI Nets (Li, Abtahi & Zhu., 2017)

3.5 Hypotheses and analytical experiments

As we explained in the introduction of this chapter, the goal of this work was both to classify the emotion using each of the four modalities and to explore the unique characteristics of the EEG modality. Although the former is a very important topic, it has been explored by many researchers, therefore the latter will be the focus of most of the analyses in this section.

Before describing the details of each analysis, we first describe the experimental setup for the models. The following configurations were shared among all the analyses unless otherwise is explicitly stated.

Even though LSTM and DBN are both deep models and are able to extract hidden information from the data and represent it as learned features, each has different strengths. Since each modality captured in our data is a sequence of information over time, we used the LSTM, which is very powerful in dealing with temporal information. On the other hand, DBN is a very powerful feature extractor when pre-trained properly. For this reason, either LSTM or DBN is more suitable depending on the task. These models have also been combined and used as a hybrid to provide strengths from both models (Giri et al., 2016). In this paper, we will focus on the comparison of LSTM and DBN for all modalities instead of integration.

3.5.1 Data preparation

In this section, we provide details and parameter configurations for the different feature extraction methods.

- ***Images:***

For each of the 24 images, we extract a feature vector size of 2048. When using LSTM, these feature vectors were provided to the model as a sequence and an output was reported after the entire sequence. Given a sequence of n frames $Xi \in \{X1, ..., Xn\}$, the target prediction is the class

of the last *Xn* frame (Figure 24). When the length of the sequence of images was shorter than what the model expects, we padded the sequence with a blank black frame. In such cases, the sequence of images was padded at the beginning with black images and the features extracted from those frames were also appended to the rest of the feature vectors. On the other hand, when the data was provided to the DBN, all feature vectors from the sequence were concatenated and used as a single input vector to the model. Note that each image frame has been turned into a feature vector of a length of 2048 using the ROI Net, so the input dimension to DBN from 24 frames is 24*2048.

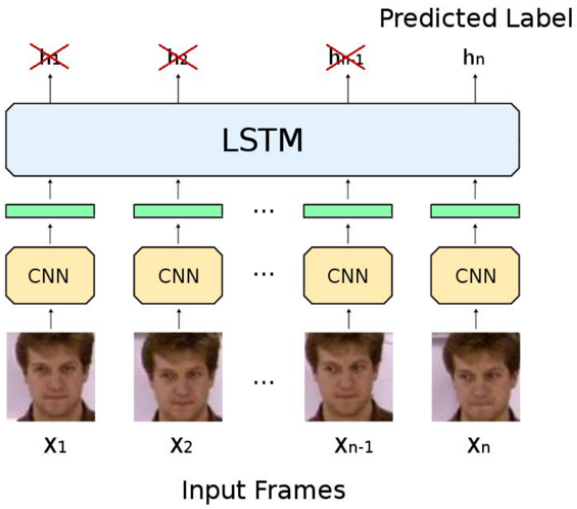

Figure 24. Feature extraction via CNN and prediction using LSTM (Bellantonio, 2016)

- ***Voice:***

The processing of voice has a few significant differences with chapter 2. Here, we use a Hamming window of size 20ms with a 10ms offset, following the work presented by (Wang, 2014). From each 20ms interval, we then extracted MFCC features and chose the 20 most significant coefficients. The input to the LSTM model was a sequence of such feature vectors. For the DBN, we concatenated these feature vectors into a single vector. Again, if the sequence

was shorter than expected, we padded it with a frame of silence. The MFCC features extracted from these silent frames were used normally as part of the sequence. We also used the i-vector features along with PLDA, for which we set up the input sequence similar to LSTM. For the extraction of the i-vector features and classification of the feature vectors using PLDA, we utilized the MSR Identity Toolbox (Sadjadi et al., 2013).

- ***EMG and EEG:***

Since both EMG and EEG are non-stationary signals that share similar characteristics with the voice signal, we used the same settings for processing them. The same 20ms window with 10ms offset was used to cut the signal of each of the six channels into a sequence of intervals. WT was applied to each interval channel by channel and the 20 most significant coefficients were kept as a feature vector for each channel. To apply LSTM to EMG, the WT coefficients of all 6 channels from a single interval were concatenated and used as the feature vector associated with that time step. Similarly, the WT coefficients of all 8 channels of EEG were concatenated to form the feature vector of each time step. The input feature vector to DBN on the other hand, was formed by concatenating the WT coefficients of the entire sequence. In the case of EEG, shorter sequences were padded with WT coefficients corresponding to Electrocerebral inactivity (ECI) or electrocerebral silence (ECS), which is defined as no EEG activity over 2μV (ACNS, 2006). In this case we padded the sequence with zeros to represent the inactivity of the area, both for EMG and EEG, and extracted WT features from the padded sequence.

3.5.2 Artifact removal

EEG recordings are usually corrupted by spurious extracerebral artifacts, which should be rejected or cleaned up. Since manual screening of human EEGs is inherently error prone and might induce experimental bias, automatic artifact detection is extremely important and is the

best guarantee for clean results. The impact of muscular activity on the signal can be evaluated using artifact removal approaches by placing emphasis on the analysis of EEG activity (Vialatte et al., 2008; Hu et al., 2015). To remove the effect of EMG from EEG signals, we use the AAR plug-in for EEGLAB (Gómez-Herrero et al., 2006). We will compare the results on unfiltered EEG data and compare it with filtered EEG, but once the positive effect of the artifact removal procedure is experimentally proven, we will continue the rest of the analyses with filtered EEG signals only.

3.5.3 Initial experiments

We began our analyses by comparing different modalities for emotion recognition using DBN and LSTM as the classification methods. First, we randomly chose 60% of the samples for training and the remaining 40% for test. For each modality, the entire 5-second sequence was provided to the model and the model classified the sequence into one of the seven emotions. For image sequences, a more accurate term to use would be facial expressions instead of emotions, but since our goal was to label the data based on the underlying emotion, we simply refer to the task as emotion recognition/classification.

The process of randomly dividing the data into 60% training and 40% test samples was repeated 100 times and the results averaged over all repetitions. The goal was to classify the emotions, not the subjects. The training and test datasets do not overlap from an emotion classification perspective, but they did contain samples from the same subject expressing the same emotion on different trials. Emotions were quite different across sub-sessions, and the order of the emotions that were being expressed was randomized and different across the sub-sessions to minimize information leak. Figure 25 summarizes the means and standard deviations of the results, showing that LSTM (the discriminative model) classifies better on voice and

images, whereas DBN (the generative model) performs better for EMG and EEG data. The results also show that facial images discriminate between emotions the best, followed by voice, EMG, and then EEG activity.

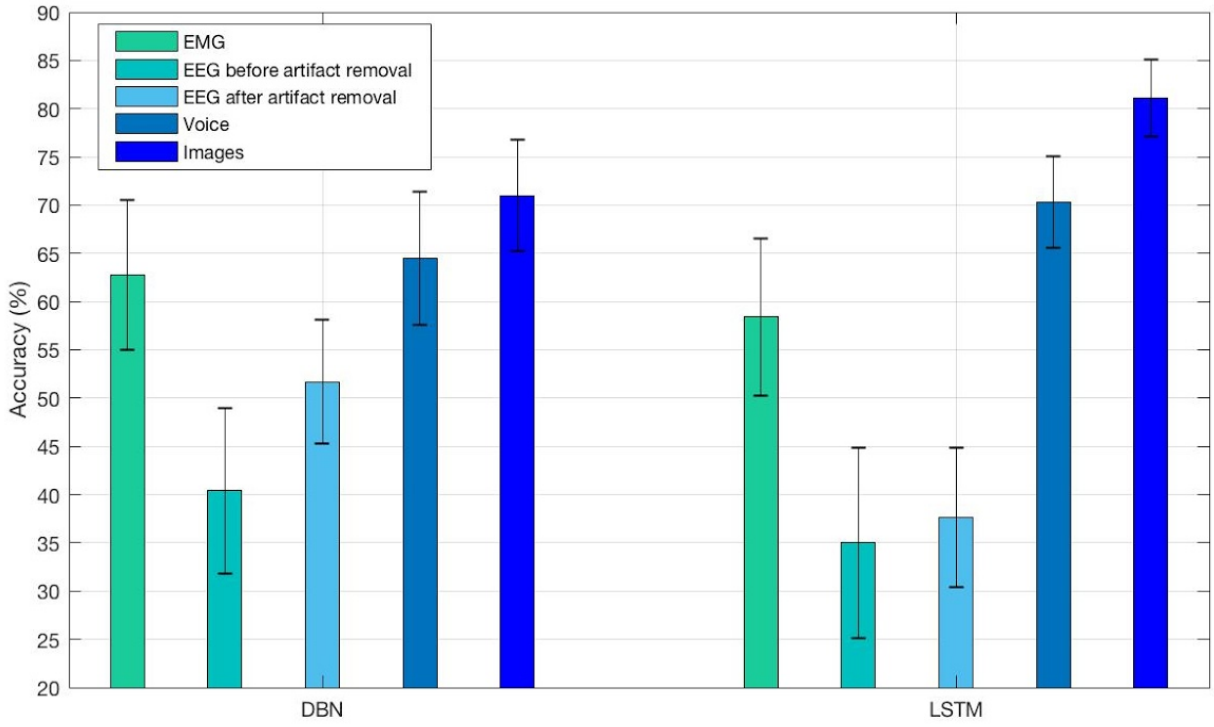

Figure 25. Comparison of emotion recognition accuracies on all modalities using DBN and LSTM

As also seen in Figure 25, the EEG signals result in higher accuracies after artifact rejection. However, the classifier, especially in the case of EEG signals, are not as accurate and accuracy variations across the trials was high. Nonetheless, classification performance was still much greater than would be expected by chance (14.3%). The confusion matrices in Table 1 and Table 2 shows the class assignments for classifying EEG after artifact removal with DBN (Table 1) and images with LSTM (Table 2). As can be seen in this table, the model gets easily confused between different classes, especially for Sad (as Disgust, Fear and Anger for around 10%) and Fear (as Disgust and Anger for around 10%). We also note that approximately 10% of each of

the non-Neutral emotions (except Surprise) was classified as Neutral, which indicates that the overall emotion information provided through the EEG signals was not very strong. The images still do a much better job in classifying the emotions; only Sad was classified as Neutral for more than 10% of the classifications.

The fact that several of the examples were classified incorrectly led us to manually investigate the data and to check if classes with different labels have similarities. We suspected that the emotions do not exactly start or end in the dedicated time slot and might leak into the previous or following slots. For EEG signals, this can particularly be more explainable due to the possible delay in emotional state taking effect in the brain (Zhang et al., 2013), and in more general visual classification tasks (Spampinato et al., 2016). In the next analysis, we verify this hypothesis by analyzing different parts of each sequence separately. In the next analysis, we verify this hypothesis by analyzing different parts of each sequence separately.

Table 1. Confusion matrix of DBN classifier on filtered EEG

	Neutral	Sad	Happy	Disgust	Fear	Surprised	Anger
Neutral	49.8%	10.3%	0.7%	4.7%	5.1%	19.7%	10.0%
Sad	14.9%	45.6%	0.1%	9.6%	9.7%	5.5%	14.3%
Happy	9.6%	5.6%	59.8%	0.1%	5.1%	10.3%	9.9%
Disgust	10.3%	5.2%	0.0%	70.6%	4.9%	0.3%	9.7%
Fear	10.2%	5.1%	5.0%	9.9%	48.6%	5.4%	14.9%
Surprised	0.2%	0.5%	10.3%	4.5%	10.1%	64.6%	9.8%
Anger	9.7%	4.9%	0.5%	5.3%	4.9%	4.7%	69.3%

Table 2. Confusion matrix of the LSTM classifier on image sequences

	Neutral	Sad	Happy	Disgust	Fear	Surprised	Anger
Neutral	69.2%	15.1%	0.1%	0.2%	9.9%	0.1%	5.4%
Sad	15.1%	79%	0.2%	0.3%	5.4%	0.4%	0.1%
Happy	0.1%	0.2%	88.1%	0.3%	0.2%	10.3%	0.2%
Disgust	5.1%	0.2%	0.4%	84.1%	0.3%	0.4%	10%
Fear	0.2%	5%	0.3%	0.4%	78.9%	15%	0.1%
Surprised	0.3%	0.1%	5.3%	0.1%	10.3%	83.9%	0%
Anger	5.3%	5.1%	0.2%	10.3%	0%	0.2%	78.9%

3.5.4 Dividing the data into more meaningful segments

The easiest way to divide each sequence into sensible intervals is by using the beginning and end of the voice signal to mark the sequence. Since we previously found that the maximum length of the utterance was approximately 2.5 seconds but the data was sampled for 5 seconds after the emotion word onset, we segmented each sample into three parts: *pre-speech*, *during-speech*, and *post-speech*, using the beginning and end of the voice signal as the timestamps to divide the sample. “During-speech” starts as soon as the subject begins uttering the sentence and ends once the utterance ends. This is done by automated speech segmentation. We standardized the length of all “during-speech” segments (by resampling) across the entire dataset so that they were all 2.5 secs. We considered the 1.25 sec segment before the beginning of speech as “pre-speech”. The 1.25 sec segment beginning at the end of voice was considered as “post-speech”.

Using this segmentation method, we performed a series of analyses where we compared every segment against all the three segments from the same emotion across trials. We repeated this analysis for all modalities, using both DBN and LSTM, for a total of 9x2 results for each

modality (except speech). Since i-vector is often used in speech signal classification, we compared the classifiers with the i-vector approach as well. In addition, since we used voice to split the segments, pre-speech and post-speech segments are not meaningful for voice-based emotion recognition; thus, we did not include those combinations in our analyses.

Note that, for instance, when we test post-speech against pre-speech, we randomly chose 60% of the pre-speech segments as the training set and the post-speech segment part of the remaining 40% as the test set. This process was repeated 100 times and the accuracies were averaged. On the other hand, for post-speech against post-speech (or any other matching pair), we randomly chose 60% of the segments for training and the remaining 40% for test, as usual.

Furthermore, since the length of the pre- and post-speech sequences are shorter than during-speech, we padded the shorter sequences to the length of the longer sequence in order to test and train on a non-matching pair (e.g. pre- against during-speech).

Figure 26 through Figure 29 demonstrate the accuracies of DBN and LSTM classifiers on all modalities, per segment. We compared every segment of an emotion against other segments (including the segment itself) of the same emotion to verify whether the emotion consistently continued over the entire 5 second interval. The following observations should be noted from these results:

1) The overall observation based on Figure 26 through Figure 29 is that even though all segments that are compared belong to the same emotion, they do not exactly match if the pair is from non-matching segments, i.e. any combination other than pre-vs-pre, during-vs-during and post-vs-post-speech. This observation holds for all modalities. In particular, the large difference between the pre and post-speech convinced us that the data is contaminated before and after the speech, either by random emotions/facial

expressions, or hypothetically by the leakage of each emotion into the following, which result in mismatch between the pre and post- speech segments.

2) Another important observation is that post-vs-post-speech comparison is always more accurate than pre-vs-pre-speech for all modalities. By manually examining the data, we realized that the subjects tend to keep the same facial expression after the 5-sec duration of the trial, until they fully read and process the label displayed on the next trial and then switch to the next emotion. This makes the post-speech segment more stable compared to the pre-speech segment. We suspect that the same phenomenon happens with EEG signals, similar to the observation made in (Spampinato et al., 2016; Palazzo et al., 2017). To further test this hypothesis, we performed a more thorough analysis. The results will be reported in Sections 3.5.5 and 3.5.6.
3) Interestingly, the post-speech segment is more accurate in classifying the emotions compared to during-speech for EEG signals. This suggests that the EEG response begins taking place slightly later than other modalities and stays active longer or that the movement contaminated EEG signals are not as reliable. We will investigate this more thoroughly later in this section.
4) For EMG and EEG signals, DBN often does a better job in classifying the emotions correctly. In contrast, LSTM performs better on image sequences. This observation holds for results on both the whole segment and sub-segments. This can be due to generative vs. discriminative capabilities of the models. EEG and EMG require a model with a strong ability to extract hidden information within the data. However, the image sequences and voice signals can readily be classified using LSTM, especially since these images and sounds have already been processed by another deep model, the CNN+ROI

platform and MFCC feature extractor, respectively, and valuable information has already been extracted from the data before the LSTM was applied.

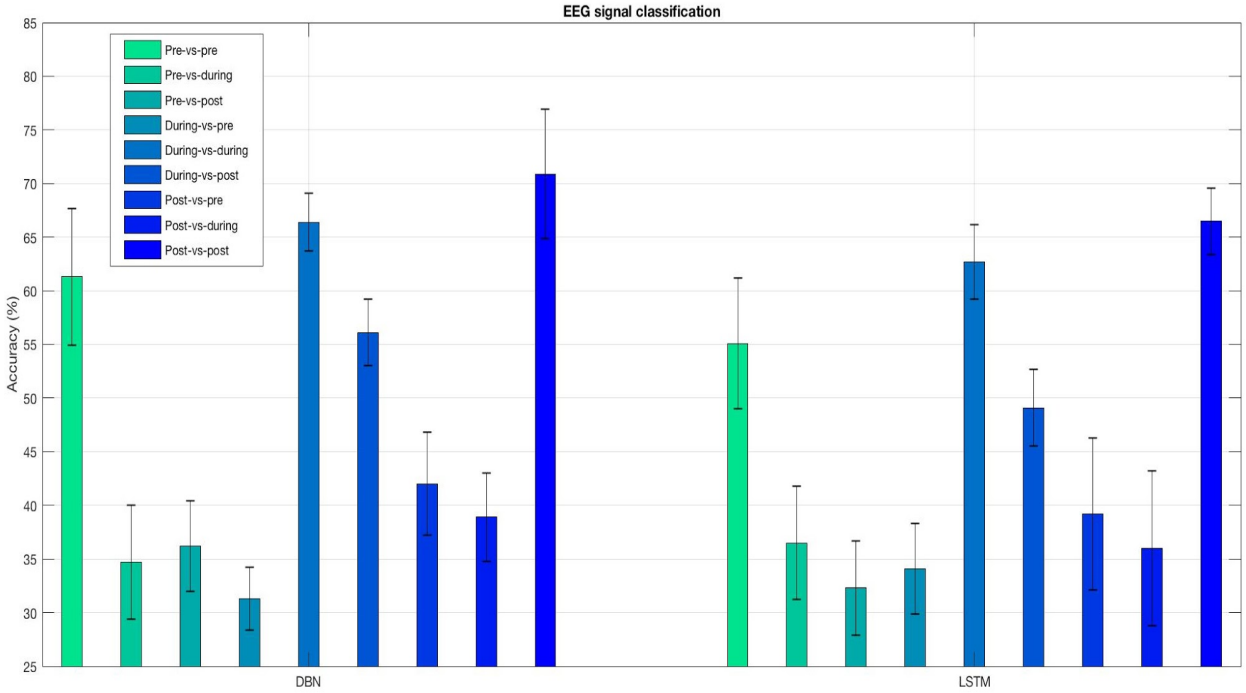

Figure 26. Classification of emotions based on EEG signals

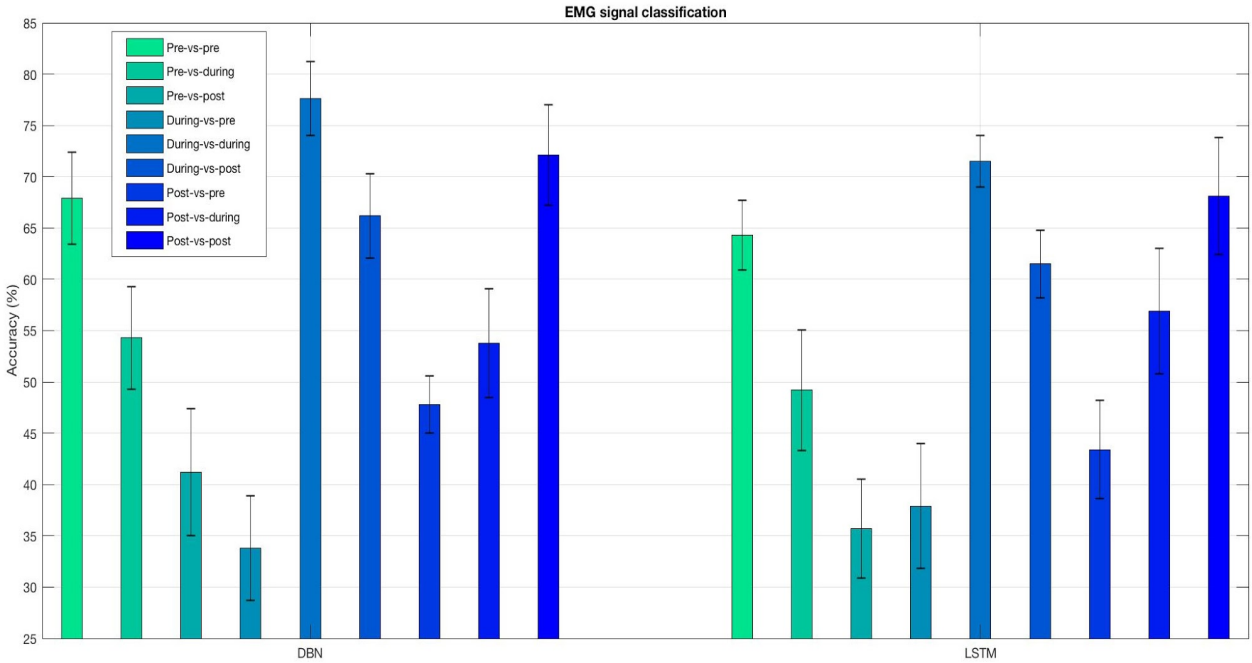

Figure 27. Classification of emotions based on EMG signals

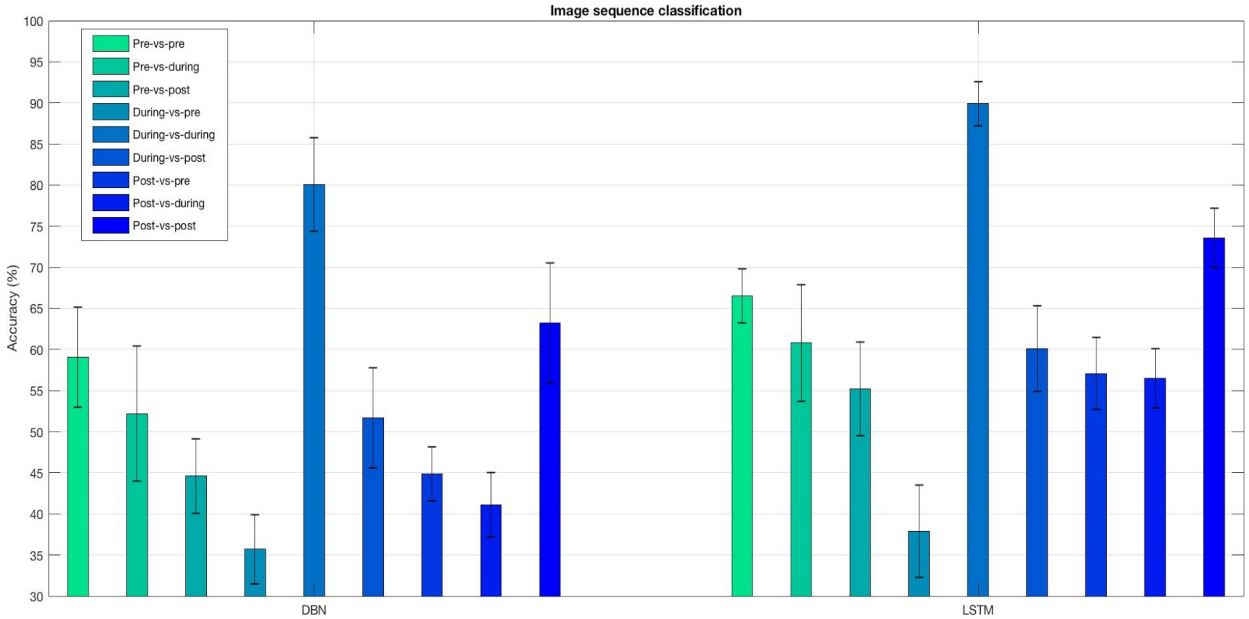

Figure 28. Classification of emotions based on image sequences

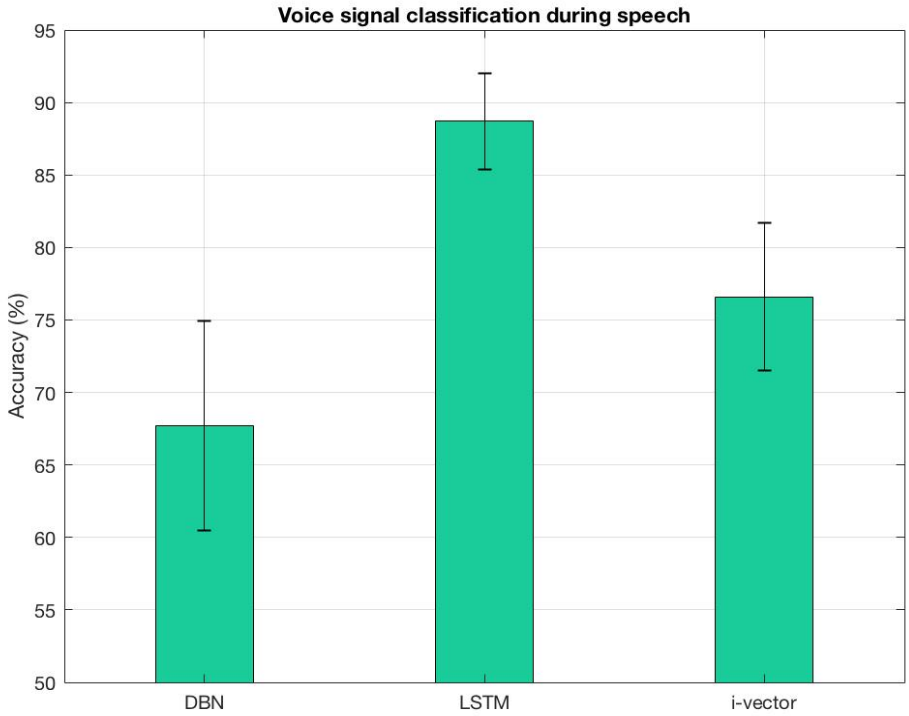

Figure 29. Classification of emotions based on voice signals

3.5.5 Unclear boundaries between consecutive emotions

Based on the previous observations (inaccuracy of the comparisons between pairs of different segments), we trained the models on post-speech from the current expression and tested if this emotion can be detected in the pre-speech signal from the next expression. We applied LSTM on image sequences, and DBN on the EEG and EMG signals, since they have shown the best performance on those modalities respectively (Figure 30). As can be seen in Figure 30, the post-speech of the current emotion signals and the pre-speech of the following emotion signals match with a surprisingly high accuracy for the EEG data (62.8%). We repeated this same analysis except we switched the training and test sets, i.e., we trained the models on pre-speech from the next emotion (using the current emotion as the label) and tested them on the post-speech from current emotion. The accuracy in this case was 66.1%, which is very close to (actually higher than) the accuracy in our previous analysis (62.8%). This is not the case for EMG or images, with fairly low classification performance.

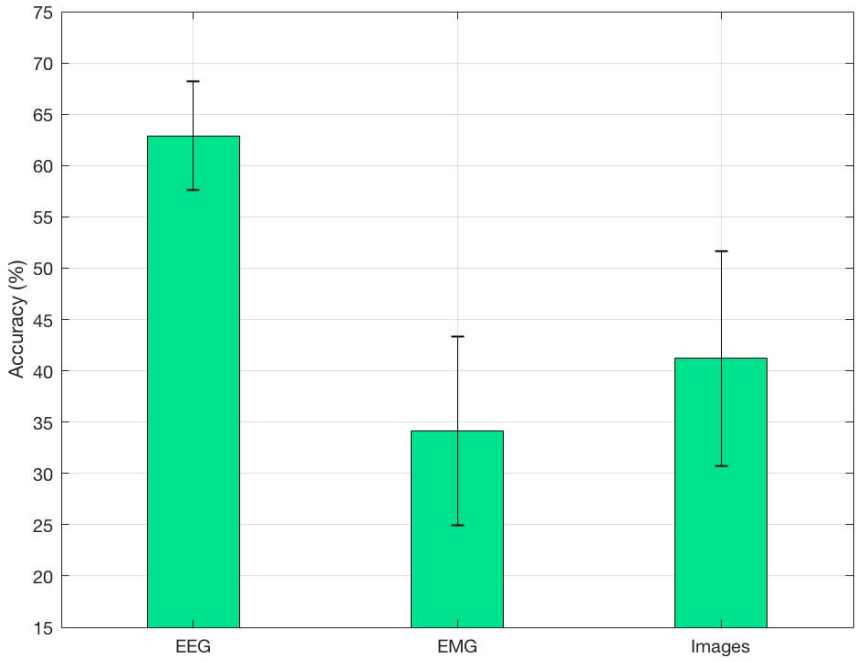

Figure 30. The aftereffect of EEG compared to EMG and Images

3.5.6 Continuation of emotions through time

The EEG aftereffects can be controlled for by giving the subjects enough time to recover from the emotions, as in (Palazzo et al., 2017; Spampinato et al., 2016). In those studies, the subjects were shown a sequence of images for 25 secs while EEG activity was recorded, followed by a 10 sec pause where a black image was shown. The black image was used to "flush" any high-level class information present from the previous one. We, on the other hand, analyzed the data in order to check the length of this aftereffect by comparing each emotion trial with the next five trials. We performed this analysis for EEG (Figure 31), as well as EMG and images (Figure 32), to show that unlike other modalities, this effect is unique to EEG signals. Similar to the previous analysis, we applied LSTM on image sequences and DBN on EEG and EMG.

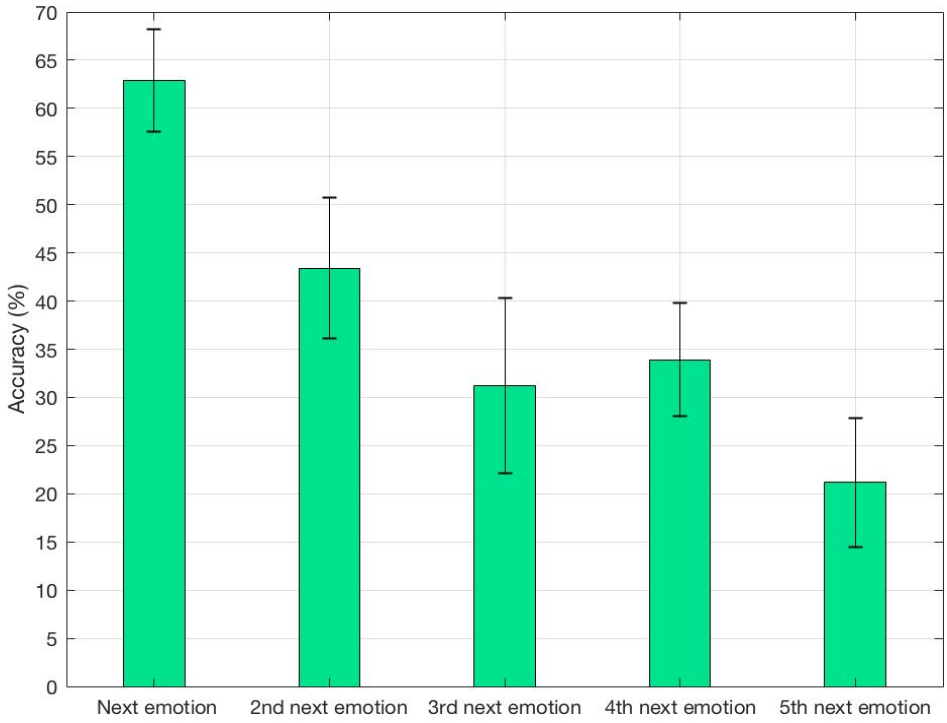

Figure 31. The aftereffect of EEG propagated through the next five emotions

Figure 31 shows that the emotion aftereffect is the strongest into the next trial (within 10 seconds), and still has some effect in the n+2 trial (within 15 seconds), but gradually decreases after the n+3 trial. Also, Figure 32 shows that unlike EEG, the EMG and image modalities reflecting a given emotion do not significantly propagate through the next trials and their effect only lasts through the pre-speech segment of the emotion immediately following the current one. Again, we do not track the effect of audio in this case, since the audio signal does not appear throughout the pre-and post-speech segments.

We should note that the pure random chance of each emotion is around 14.3% (1 in 7 emotions) and the results we obtained for EMG and images after the immediate next trial are close to chance and only slightly higher. Note that the probability of the same emotion appearing in the sequence in each of the next 2nd, 3rd, … trials was also 14.3%.

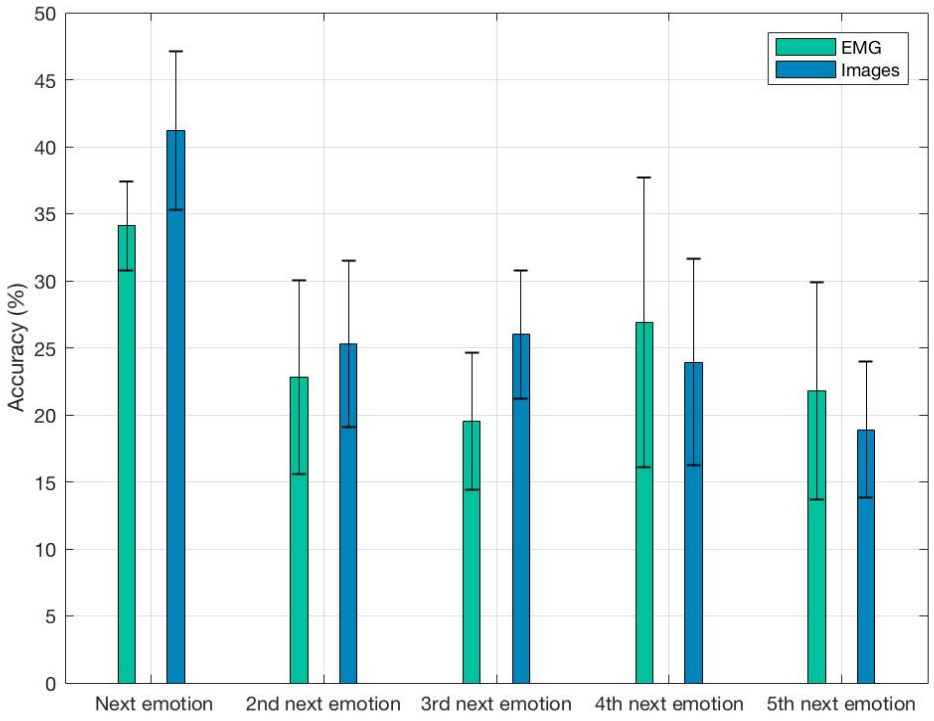

Figure 32. The aftereffect of EMG and images throughout the next five emotions

3.6 Conclusions

This chapter presents a thorough study on emotion recognition using four different modalities – audio, video, EMG and EEG. To this end, we collected a dataset with 7 emotion categories, the 4 modalities, and 12 human actor subjects. Both generative models (DBNs) and discriminative models (LSTMs) were applied to the four modalities. Our analyses indicate that LSTM is better for classifying information from audio and video, each with their own sophisticated feature extractors (MFCC and CNN), whereas DBN is better for classifying information from both EMG and EEG.

In addition to the conclusions we can make on characteristics of different models given the results of our analyses, we made interesting observations on the data modalities and their distinct properties. One interesting finding is that the EEG signal lasts much longer that the other modalities and the brain activities corresponding to the emotion can get dragged through time, which is a significant difference between EEG and other modalities that we measured in our work.

We examined how different stages of a trial (pre-speech, during-speech and post-speech) and the following trials affect EEG signals and found long-lasting neural signatures that represent different emotional states. This makes the post-speech segment more accurate in classifying the emotions compared to during and post-speech for EEG signals, as opposed to other modalities.

We believe that the dataset collected in this work can be valuable for affective computing and facial analysis, thus it will be made publically available following publication. This paper has focused on the comparison of the four modalities, especially the two bio-sensing datasets (EEG and EMG) versus the commonly used visual and audio data. In particular, one of the most interesting aspects of the analyses is the observation that neural signals conveying an emotion are

long-lasting and can be detected by the use of machine learning. This kind of temporal effect has been noted in the psychology and neuroscience literature, but this seems to be the first time it has been exploited by the computer vision community in such a significant capacity.

In the future, we would like to further this research in three directions. The first is to integrate the modalities to optimize performance by using the results of this comparison study. This can be done either at the feature level (early fusion) or the classification level (late fusion). In particular, since LSTM works better on audio and video, and DBN works better on EEG and EMG, it would be interesting to develop models combining generative and discriminative neural networks as in (Giri et al., 2016), but for emotion recognition. We would like to implement and compare both approaches. The second is to compare machine algorithms and humans in reading the emotion from audio and video, drawing more insights into emotion recognition. The third is to investigate how sensing processing can be improved (especially on EEG and EMG) to obtain more robust signals for reading emotions.

Chapter 4

Conclusions, Discussions and Future Work

4.1 Conclusions and Discussions

4.1.1. Summary of the Thesis

As a summary, the work presented in this thesis can be divided into two main parts: 1) multimodal data integration for speaker recognition and identification, and 2) analysis of different modalities and their associations for emotion recognition. The major highlights of our work include the following five aspects:

- ***Models.*** Choosing the most suitable machine learning models is the first important thing, including deep learning models that fall into the categories of generative and discriminative methods, as well as baseline methods against which we compare our deep learning models. This includes determining the best models and the best settings for model parameters through both the understanding of the state-of-the-art and best practices, and the thorough experimental studies in detailed comparisons.
- ***Dataset collection and cleaning.*** Two datasets have been collected throughout this research, for speaker recognition and emotion analysis respectively. Both of them have multimodal sensory data, with the first one having audio, video and EMG, and the second having audio,

video, EMG and EEG. We believe that these datasets can be very beneficial to the community. These datasets will soon become publicly available for research purposes.

- ***Features.*** Selecting the types of information to extract from different modalities that are the best for the given tasks was a challenge. We have used both traditional and well-studied feature extractors and state of the art deep learning models for feature extraction and transfer learning. The right combination of these features proved to be powerful in both applications: whenever the data is well-studied and well-understood, specifically-designed feature extractors, such as wavelets, MFCC, lip contours, are very effective; on the other hand, deep learning methods seem to win for those less-studied and yet complicated data such as images for facial expression.
- ***Audio, video and biomedical modalities and their relationship.*** We are not only interested in combining/integrating the modalities for better performance, but we also explore the distinct characteristics of each modality and the correct way to interpret and utilize them. This was accomplished through extensive analysis of modalities and breaking the data into segments for more detailed investigation into the behaviors of sensory modalities in time (as for EEG with pre- during- and post-segments), frequencies (as for EMG in using wavelets and audio in using MFCC) and space (as for facial images using Regions of Interest (ROI)).
- ***Multimodality.*** Training and testing the model on all modalities was expected to improve the classification performance, as proved by our analyses. These includes: the integration of multimodalities and the substitution of one modality by the other. The trained model in the speaker recognition case has the potential to be tested on data with missing modalities.

4.1.2. Further Discussions

We would like to further discuss some specific pitfalls and lessons learned in three important aspects: the biomedical sensors, the multimodal features, and the deep learning models.

1) The Biomedical Sensors

The analyses we have conducted can be divided into two parts according to the goals of the analyses with respect to biomedical sensory data: The use of EMG to determine the person's identity, and the application of EEG/EMG to determine how the person feels.

In the first group of analyses, we focused on EMG and its capabilities in determining "who you are", when compared/combined with audio and visual information. The significant finding is that the EMG modality boosts the speaker recognition rate and is even capable of replacing or complementing the audio modality, if audio is noisy or unavailable.

In the second group of analyses, we focus of biomedical sensory information to determine "how you feel". This includes the correlation between EMG and EEG data, which was to some extent solved using artifact removal techniques. The experiments in this part of our work are mainly focus on EEG, it's characteristics, and its effectiveness in detecting the emotional state in comparison with visual, audio and EMG modalities. EEG might not be the most powerful means to determine emotions, but it has very interesting characteristics that are very unique to this modality.

Previous studies prove that the brain areas corresponding to emotions do not exactly get activated when the facial and vocal expressions appear. These activities also do not exactly last as long as the facial and vocal expressions do. But few through studies with sensory measures has been performed. Our analysis experimentally confirm these properties. We have shown that probably because the emotions we capture in this study are acted, it takes the corresponding

brain areas some time to get activated. In addition, we have also found that brain activities get dragged through time. In other words, while the person is done with acting a certain emotion (or it is better to use the term facial expression in this case), the brain still feels that emotion until the electrical activities of those areas in the brain vanish. This causes leakage between consecutive emotions in our experiments. In order to harness this property, we divided the EEG signals into 3 pieces: before, during and after the emotion is acted. For all other modalities (visual and vocal), the highest accuracy in emotion recognition is achieved while it is being imitated. For EEG on the other hand, the most effective part of the EEG signal is the part that is captured after the emotion is acted. This phenomenon results in leakage between consecutive emotions, but at the same time, explains the lower classification accuracies using EEG compared to other modalities. This might indicate two things: (1) the brain signals are different from other human-perceivable signals like audio and video and even EMG (muscle movements); (2) The EEG sensors are still not robust enough to obtain the best emotion states in the brain.

2) The Multimodal Features

We apply the most appropriate signal transformation or feature extraction for each modality of sensory data we can possibly obtain before fed into the learning models in chapters 2 and 3.

Both EMG and EEG are filtered and transformed into frequency domain using wavelet transform with bandwidth selection, which has shown better performance than directly using their raw signals in our early analysis; some of the results are included in the thesis. The dimensionality of the wavelet coefficients extracted from each channel is 66 in our experiments after some initial tests with various sizes . These coefficients are concatenated for all 8 EEG and 6 EMG channels separately, to form the EEG and EMG feature vectors.

On the other hand, voice and images go through feature extractors that are specifically designed for those modalities. For instance, a pre-trained CNN with ROI functions that is trained on the well-studied facial action units is used to extract features from the images, and the human-perception-based MFCC features are extracted from audio signals, so the resulting feature vector of each case is already adapted to the task: the dimension of the CNN feature for each image is 2048, whereas the dimensionality of the MFCC feature of a 2.5 second audio clip is ((2,500/10)-1)*20 = 4,980, as explained in section 3.4.2. We also have the sequence of 20 key-points extracted from the mouth area for lip motion in Chapter 2, where the coordinates of these points are concatenated and used as visual information for speaker identification. This seemed to be enough for our experimental setup, although we can extend it to an image of mouth area or even the entire face in our future work.

3) The Deep Learning Models

In the analyses performed in Chapters 2 and 3, each type of models had a different behavior when applied to each of the data modalities. The baseline methods are chosen from the most relevant literature on each modality. Random Forest and SVM are picked as out of the box multi-purpose models that have been used in multimodal data integration and classification. GMM and i-vector features in combination with Linear Discriminant Analysis are chosen since they are the state of the art voice processing methods. In the majority of our experiments, the deep learning models outperformed the baseline methods.

For deep learning models, we have found in Chapter 3 that DBN worked better on EMG and EEG modalities, while LSTM performed better on image sequences and voice. Here are some more detailed analysis.In case of EEG and EMG, the transformed features (wavelet coefficients) are still low-level features, therefore using DBN proven to be a powerful tool to extract higher-

level features from them, as it is a generative model. Thus, DBN is responsible both for feature extraction and classification. For voice and images though, LSTM can discriminate the extracted feature vectors faster and easier than a DBN, since the hidden information is already extracted from the data and this information is sufficient enough for the LSTM to perform the classification as a discriminative model.

Apart from the two categories of deep learning models (generative and discriminative), a third category can also be defined, which consists of "hybrid" deep learning models. Giri et al. (2016) performed a thorough research on combination of DBNs and LSTMs to benefit from advantages of both type of models. The discriminative models are architectures that have direct ability to classify. A few examples of discriminative architectures, for instance, are CNN, recurrent neural network (RNN), or LSTM. On the other hand, even though not usually used to classify in a direct way, the generative models are very handy for the classification and regression tasks, especially in the stages of data preparation, such as initialization process and pre-training for the training parameters. An example of generative models is DBN, where the model is first pre-trained on the data without taking into account the class labels or target values. The model is then refined using the information provided by the targets to turn into a classifier/regressor.

In summary, the classification of all of our modalities can be viewed as a time series classification problem. In the future, one of our primary goals will be to measure and evaluate the performance of deep hybrid architectures and compare them to the individual generative and discriminative models that we have used in our work, as well as the state of the art classification approaches applied to the emotion recognition problem in the literature.

4.2. Limitations and Future Work

4.2.1. Limitations

There are a few limitations that we faced while conducting the analyses, especially for the biosensor data (EEG/EMG), including:

1) The multi-channel data use:

As we explained in Chapters 2 and 3, we concatenate the sensor readings from all EMG/EEG channels in the analyses to form a single vector and feed it to the model. This approach usually works best if the number of samples in the dataset is large. In that case, the model would still be able to determine the cutting point between the channels in speaker identification case, and the temporal relationship between pieces of the vector in emotion recognition analyses. Since our dataset is relatively small, concatenation might not be the best approach. One way to solve this problem is to train one sub-model per channel and combine the resulting features through a shared layer; the same way different modalities are combined in out models. This approach will significantly reduce the number of model parameters that need to be tuned during training and as a result, reduces the required training time.

2) The parameters of the deep learning models and the models themselves:

Although the models we have used in our analyses have been selected based on related literature, there might be some limitations in a few cases. For instance, DBN is not normally the best choice when temporal information is involved. Specifically, as explained above, we had to concatenate signals from consecutive timesteps in order to train the DBN, which is not the most effective method in this case. Also there are more baseline methods that we can compare our study with. For the deep learning models, we can always perform a parameter search to fully optimize the model for the task. Currently, the parameter values (number of layers, number of

units per layer, etc.) are selected from a set of values, but we can extend this to approaches such as random search, grid search, Bayesian optimization, or other parameter optimization algorithms.

3) The integration approaches:

There are mainly two fusion approaches used in the literature to integrate modalities of multimodal data: early (feature level) and late (decision level) fusion. Methods that rely on early fusion first extract unimodal features. After analysis of the various unimodal streams, the extracted features are combined into a single representation. After combination of unimodal features in a multimodal representation, early fusion methods rely on supervised learning to classify the samples. In our case, we used concatenation of unimodal feature vectors from the EMG and EEG channels to obtain a fused representation of these modalities. Other modalities are also fused using a shared representation layer in the models.

The approaches that rely on late fusion also start with extraction of unimodal features. In contrast to early fusion, where features are then combined into a multimodal representation, approaches for late fusion extract information directly from unimodal features. As an example, separate generative probabilistic models can be learned for the visual, voice, EMG and EEG modalities. The scores obtained from these models are combined afterwards to yield a final detection score.

The early fusion is advantageous in that it can utilize the correlation between multiple features from different modalities at an early stage which helps in better task accomplishment. Also, it requires only one learning phase on the combined feature vector. However, in this approach it is hard to represent the time synchronization between the multimodal features. This is because the features from different but closely coupled modalities could be extracted at

different times. Moreover, the features to be fused should be represented in the same format before fusion. In addition, the increase in the number of modalities makes it difficult to learn the cross-correlation among the heterogeneous features.

The late fusion strategy has many advantages over feature fusion. For instance, unlike feature level fusion, where the features from different modalities (e.g. audio and video) may have different representations, the decisions (at the semantic level) usually have the same representation. Therefore, the fusion of decisions becomes easier. Moreover, the decision level fusion strategy offers scalability (i.e. graceful upgradation or degradation) in terms of the modalities used in the fusion process, which is difficult to achieve in the feature level fusion. Another advantage of late fusion strategy is that it allows us to use the most suitable methods for analyzing each single modality. This provides much more flexibility than the early fusion. On the other hand, the disadvantage of the late fusion approach lies in its failure to utilize the feature level correlation among modalities. Moreover, as different classifiers are used to obtain the local decisions, the learning process for them becomes tedious and time-consuming.

To exploit the advantages of both the feature level and the decision level fusion strategies, several researchers have opted to use a hybrid fusion strategy, which is a combination of both feature and decision level strategies. We are interested in further studying these three approaches and selecting the most suitable strategy for our tasks.

4.2.2. Future Directions

There are multiple future directions that we would like to further explore:

1) First, as we mentioned earlier, the deep learning models that we implemented are able to work in absence of one of the modalities. This is particularly of interest for the EMG and EEG signals, as they are used to capture additional and/or inherent features of

speakers/emotions but it is not feasible to attach those types of sensors to the users in daily practice. So, tentatively the model will be trained on the data collected in lab settings, which includes EMG/EEG signals, but will be tested on only other modalities that are easily available in practice.

The high-level method for filling in the missing modalities has three main steps:

a) Building a joint density on all modalities,

b) using states of top-level hidden units as joint representation, and

c) sampling from the conditional density to fill in the missing modality.

2) The second suggestion is to combine generative and discriminative models for multimodal emotion analysis. We realized that each model is most suitable for specific types of modalities. Discriminative models are powerful in extracting hidden information in the data, while discriminative models can successfully separate the extracted features into meaningful categories. The combination of these two types of models could result in much more powerful classifiers. One of the improvements that we will apply in our future work is to combine DBN with LSTM, similar to the work of Giri et al. (2016). An idea for our future research direction is to fuse generative ability of the DBN to extract multi-level hierarchical features and determine the final class label using the time series discrimination capability of LSTM.

3) Another interesting future work will be to generate modalities given the other ones. This is particularly useful if we need to convey information to a disabled individual, who lacks a certain sense or the ability to show feedback in a certain way, but information can be transferred to, or captured from them through other senses/reactions. This is similar to dealing with missing modalities and the multimodal models that are able to learn a shared

representation of the multimodal data are capable of performing such modality generation task. The results would be useful in particular in assistive devices.

4) Multi-channel data use and fusion strategies are also two other directions that we would like to explore. In other words, we would like to address the limitations of our methods in our future work. We explained the details of these limitations in section 4.2.1. We currently concatenate the multi-channel data, which limits the ability of the model to distinguish the channels and the underlying temporal information within the data. We will test other methods specifically to integrate channels of EMG and EEG signals. Same idea applies to the integration of different modalities with early, late and hybrid fusion strategies.
5) The goal of wavelet analysis is to decompose signals into several frequency bands. When using wavelets on EEG data, it is more typical to extract the power at each frequency over time, rather than to use the coefficients. This is also one of the improvements that we would like to try in our future work.

List of the Candidate's Publications

1. Abtahi, F. Ro, T., Li, W. & Zhu, Z. (2018). Emotion Analysis Using Audio/Video, EMG and EEG: A Dataset and Comparison Study. IEEE Winter Conference on Application of Computer Vision, (WACV 2018).
2. Abtahi, F., Li, W., Zhu, Z. & Ro, T. (2017). Using EMG to Identify Speakers: A Comparison Study to Replace or Enhance Audiovisual Data. Submitted to Machine Vision and Applications (under review).
3. Li, W., Abtahi, F., Zhu, Z. & Yin, L. (Nov. 2017). EAC-Net: Deep Nets with Enhancing and Cropping for Facial Action Unit Detection. Special Issue on Computational Face, IEEE Transactions on Pattern Analysis and Machine Intelligence. Volume PP, Issue 99, 10 January, 2018 (DOI: 10.1109/TPAMI.2018.2791608).
4. Li, W., Abtahi, F., & Zhu, Z. (2017). Action Unit Detection with Region Adaptation, Multi-labeling Learning and Optimal Temporal Fusing, IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017), Honolulu, Hawaii, USA
5. Li, W., Abtahi, F., Zhu, Z. and Yin, L. (2017) EAC-Net: A Region-based Deep Enhancing and Cropping Approach for Facial Action Unit Detection. The 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017), Washington, DC.
6. Li, W., Abtahi, F., Tsangouri, C., & Zhu, Z. (2016, June). Towards An "In-The-Wild" Emotion Dataset Using a Game-Based Framework. CVPRW 2016, Las Vegas, NV.

7. Abtahi, F., Li, W., Zhu, Z., & Ro, T. (2015, June). Multimodal Speaker Recognition using Deep Belief Networks. Women in Computer Vision (WiCV) Workshop, CVPR 2015, Boston, MA.

8. Li, W., Abtahi, F., & Zhu, Z. (2015, November). A Deep Feature-based Multi-Kernel Learning Approach for Video Emotion Recognition. International Conference on Multimodal Interaction (ICMI'15), Seattle, WA.

9. Burry, A. M., & Abtahi, F. (2015). A Reinforcement Learning Approach to Character Level Segmentation of License Plate Images. U.S. Patent 14159590, Filed Nov. 2013, Published May 2015. URL

10. Abtahi, F., Zhu, Z., & Burry, A. (2015). A Deep Reinforcement Learning Approach to Character Segmentation of License Plate Images. 14th IAPR International Conference on Machine Vision Applications, Tokyo, Japan.

11. Abtahi, F., Knapp, J., & Zhu, Z. (2012, October). Using Machine Learning Techniques to Assist the Visually Impaired in Navigation and Obstacle Avoidance. 7th Annual Machine Learning Symposium, New York, NY.

12. Abtahi, F., & Fasel, I. (2011). Deep Belief Nets as Function Approximators for Reinforcement Learning (Full Paper). AAAI Workshop on Lifelong Learning from Sensorimotor Experience, San Francisco, CA.

13. Abtahi, F., & Fasel, I. (2011). Deep Belief Nets as Function Approximators for Reinforcement Learning (Short Paper). IEEE International Conference on Development and Learning and Epigenetic Robotics, Frankfurt, Germany.

14. Mafi, N., Abtahi, F., & Fasel, I. (2011). Information Theoretic Reward Shaping for Curiosity Driven Learning in POMDPs. IEEE International Conference on Development and Learning and Epigenetic Robotics, Frankfurt, Germany.

Bibliography

1. Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine learning, vol. 20, no. 3, pp. 273-297.
2. Furui, S., 1997. Section 1.7: Speaker recognition, in: Survey of the state of the art in human language technology, [Online] Available: http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html.
3. Gang, L., Hansen, J. H. L., 2014. An investigation on back-end for speaker recognition in multi-session enrollment. IEEE/ACM Trans. Audio, Speech, and Language Processing, vol. 22, no. 12, pp. 7755-7759.
4. Reynolds, D. A., 2002, An overview of automatic speaker recognition. Proc. ICASSP, Florida, 4072-4075.
5. Van Boxtel, A., 2010. Facial EMG as a tool for inferring affective states. Proc. Measuring Behavior, pp. 104-108.
6. Quan, Z., Jiang, N., Englehart, K., Hudgins, B., 2009. Improved phoneme-based myoelectric speech recognition. IEEE Trans. Biomedical Engineering, vol. 56, no. 8, pp. 2016-2023.
7. Chan, A. D. C., Englehart, K., Hudgins, B., 2006. Multiexpert automatic speech recognition using acoustic and myoelectric signals. IEEE Trans. Biomedical Engineering, vol. 53, no. 4, pp. 676-685.
8. Zhang, C., Yin, P., Rui Y., 2008. Boosting-based multimodal speaker detection for distributed meeting videos. IEEE Trans. Multimedia, vol. 10, no. 8, pp.1541-1552.

9. Hazen, T. J., Weinstein, E., Kabir, R., Park, A., Heisele, B., 2007. Multimodal face and speaker identification for mobile devices. Face Biometrics for Personal Identification. Springer Berlin Heidelberg, 123-138.

10. Çetingül, H. E., Erzin, E., Yemez, Y., Tekalp, A. M., 2006. Multimodal speaker/speech recognition using lip motion, lip texture and audio. Signal Processing, vo. 86, no. 12, pp. 3549-3558.

11. Zhang, X., Broun, C. C., 2001. Using lip features for multimodal speaker verification. at A Speaker Odyssey - The Speaker Recognition Workshop.

12. Roy, D., Shukla, A., 2013. Speaker recognition using multimodal biometric system. Proc. O-COCOSDA/CASLRE, pp. 1-7.

13. Nakagawa, S., Zhang, W., Takahashi, M., 2004. Text-independent speaker recognition by combining speaker-specific GMM with speaker adapted syllable-based HMM. Proc. ICASSP'04, vol. 1, pp. 1-81.

14. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A. Y., 2011. Multimodal deep learning. Proc. ICML-11, pp. 689-696.

15. Srivastava, N., Salakhutdinov, R., 2012. Learning representations for multimodal data with deep belief nets. Intl. Conf. on Machine Learning Workshop.

16. Srivastava, N., Salakhutdinov, R., 2012. Multimodal learning with deep boltzmann machines. J. of Machine Learning Research, vol. 15, no. 1, pp. 2949-2980.

17. Smolensky, P., 1986. Information processing in dynamical systems: Foundations of harmony theory. Parallel distributed processing: explorations in the microstructure of cognition, vol. 1, MIT Press, pp. 194-281.

18. Abtahi, F., Fasel, I., 2011. Deep belief nets as function approximators for reinforcement learning. Proc. 15th AAAI Conference on Lifelong Learning, CA, pp. 2-7.

19. Hinton, G., Osindero, S., Teh, Y. W., 2006. A fast learning algorithm for deep belief nets. Neural Computation, vol. 18, no. 7, pp. 1527-1554.

20. Erhan, D., Manzagol, P., Bengio, Y., Bengio, S., Vincent, P., 2009. The difficulty of training deep architectures and the effect of unsupervised pre-training. Proc. AISTATS, vol. 5, pp. 153-160.

21. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., 2011. Scikit-learn: Machine learning in Python. J. of Machine Learning Research, no. 12, pp. 2825-2830.

22. Muscle Sensor v3 Kit - SEN-11776 - SparkFun Electronics. [online]. Available: https://www.sparkfun.com/products/retired/11776

23. Sharma, S., Kumar, G., Kumar, S. Mohapatra, D., 2012. Techniques for Feature Extraction from EMG Signal. Intl. J. of Advanced Research in Computer Science and Software Engineering, vol. 2, no.1.

24. Phinyomark, A., Limsakul, C., Phukpattaranont, P., 2011. Application of wavelet analysis in EMG feature extraction for pattern classification. Measurement Science Review, vol. 11, no. 2, pp. 45-52.

25. Khushaba, R. N., Kodagoda, S., Lal, S., Dissanayake, G., 2011. Driver drowsiness classification using fuzzy wavelet-packet-based feature-extraction algorithm. IEEE Trans. on Biomedical Engineering, vol. 58, no. 1, pp. 121-131.

26. Wang, G., Wang, Z., Chen, W., Zhuang, J., 2006. Classification of surface EMG signals using optimal wavelet packet method based on Davies-Bouldin criterion. Medical and Biological Engineering and Computing, vol. 44, no. 10 pp. 865-872.

27. Daubechies, I., 1992. Ten lectures on wavelets. CBMS-NSF Regional Conf. Series in Applied Mathematics, vol. 61.

28. BioSemi active electrodes. http://www.biosemi.com/active_electrode.htm

29. Surface EMG Sensors, Delsys, Inc. http://www.delsys.com/products/desktop-emg/surface-emg-sensors

30. Lee, K. S., 2008. EMG-Based Speech Recognition Using Hidden Markov Models With Global Control Variables. IEEE Trans. Biomedical Engineering, vol. 55, pp. 930-940.

31. Wand, M., Schultz, T., 2011. Session-Independent EMG-based Speech Recognition. Intl. Conf. on Bio-inspired Systems and Signal Processing.

32. Breiman, L., 1996. Bagging predictors. Machine Learning, vol. 24, no. 2, pp. 123–140.

33. Schapire, R.E., Freund, Y., Bartlett, P. and Lee, W.S., 1998. Boosting the margin: A new explanation for the effectiveness of voting methods. The annals of statistics, 26(5), pp.1651-1686.

34. Schapire, R. E., 1990. The Strength of Weak Learnability. Machine Learning. Kluwer Academic Publishers, Boston, MA, vol. 5, no. 2, pp. 197–227.

35. Breiman, L., 2001. Random forests. Machine learning, 45(1), pp.5-32.

36. Friedman, J. H., 2001. Greedy function approximation: a gradient boosting machine. The Annals of Statistics, pp.1189–1232.

37. Suresh, M., Krishnamohan, P. G., Holi, M. S., 2014, Processing of Natural Signals like EMG for Person Identification using NUFB-GMM. Int. J. of Advanced Computer Research, vol. 4, no. 3, pp. 819-827.

38. Suresh, M., Krishnamohan, P. G., Holi, M. S., 2014. A over damped person identification system using EMG signal. Intl. J. of Research in Engineering and Technology, vol.3, iss. 10, pp. 54-60.

39. Suresh, M., Krishnamohan, P. G., Holi, M. S., 2011. Electromyography analysis for person identification. Intl. J of Biometrics and Bioinformatics (IJBB), vol. 5, iss. 3, pp. 172-179.

40. Hosseini, M. P., Nazem-Zadeh M. R., Mahmoudi, F., Ying, H., Soltanian-Zadeh, H., 2014. Support vector machine with nonlinear-kernel optimization for lateralization of epileptogenic hippocampus in MR images. IEEE 36th Intl. Conf. Engineering in Medicine and Biology Society, pp. 1047-1050.

41. Ho, T. K., 1995. Random decision forests. Document Analysis and Recognition. 3rd Intl. Conf. on Document Analysis and Recognition, vol. 1.

42. Asadpour, V., Towhidkhah, F., Homayounpour, M. M., 2006. Performance enhancement for audio-visual speaker identification using dynamic facial muscle model. Medical and Biological Engineering and Computing, vol. 44, no. 10, pp. 919-930.

43. Hosseini, M. P, Soltanian-Zadeh, H., Elisevich, K., Pompili, D., 2016. Cloud-based deep learning of big EEG data for epileptic seizure prediction. IEEE Global Conf. on Signal and Information Processing (GlobalSIP), Washington D. C.

44. Ioffe, S., 2006. Probabilistic linear discriminant analysis. Proc. European Conf. on Computer Vision (ECCV) 2006, pp. 531-542.

45. Kanagasundaram, A., 2014. Speaker verification using i-vector features. PhD thesis, Queensland University of Technology, Australia.

46. Reynolds, D., 2015. Gaussian mixture models. Encyclopedia of biometrics, Springer US, pp. 827-832.

47. Zeng, C., Li, Z., 2011. Application of GMM in the speaker identification system. Proc. 7th Intl. Conf. on Wireless Communications, Networking and Mobile Computing (WiCOM 2011), pp. 1-4.

48. Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. Proc. IEEE Trans. on Audio, Speech, and Language, vol. 19, issue 4, pp. 788 – 798.

49. LeCun, Y. and Bengio, Y., 1995. Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10), p.1995.

50. King, D. E., 2009. Dlib-ml: A Machine learning toolkit. J. of Machine Learning Research 10, pp. 1755-1758.

51. Barua, P., Ahmad, K., Khan, A. A. S., Sanaullah, M., 2014. Neural network based recognition of speech using MFCC features. 2014 Intl. Conf. of Informatics, Electronics & Vision (ICIEV).

52. Mohan, B. J., Babu N., R., 2014. Speech recognition using MFCC and DTW. Proc. 2014 Intl. Conf. of Advances in Electrical Engineering (ICAEE), pp. 1-4.

53. O'Mahony, M., 1986. Sensory evaluation of food: statistical methods and procedures. CRC Press.

54. Suresh, M., Krishnamohan, P., Holi, M., 2011. GMM modeling of person information from EMG signals. 2011 IEEE Recent Advances in Intelligent Computational Systems.

55. Sadjadi, S. O., Slaney, M., Heck, L., 2013. MSR Identity Toolbox v1.0: A MATLAB Toolbox for Speaker Recognition Research. Microsoft Research Technical Report.

56. Quitadamo, L. R., Cavrini, F., Sbernini, L., Riillo, F., Seri, S., Saggio, G., 2017. Journal of Neural Engineering, vol. 14, no. 1.

57. Liarokapis, M. V., Artemiadis, P. K., Kyriakopoulos, K. J., 2013. Proc. IEEE Intl. Conf. on Rehabilitation Robotics (ICORR), pp. 1-6.

58. Li, W., Abtahi, F. and Zhu, Z., 2017. Action Unit Detection with Region Adaptation, Multi-labeling Learning and Optimal Temporal Fusing. CVPR 2017. Also arXiv preprint arXiv:1704.03067.

59. Simonyan, K. and Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

60. He, K., Zhang, X., Ren, S. and Sun, J., 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

61. Katsikitis, M. ed., 2003. The human face: measurement and meaning. Springer Science & Business Media.

62. Zeng, Z., Pantic, M., Roisman, G.I. and Huang, T.S., 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE transactions on pattern analysis and machine intelligence, 31(1), pp.39-58.

63. Wang, Z., Wang, S. and Ji, Q., 2013. Capturing complex spatio-temporal relations among facial muscles for facial expression recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3422-3429).

64. King, D., 2009. dlib C++ Library-Introduction. http://dlib.net

65. He, K., Zhang, X., Ren, S. and Sun, J., 2014, September. Spatial pyramid pooling in deep convolutional networks for visual recognition. In European Conference on Computer Vision (pp. 346-361). Springer, Cham.

66. Allen, F. R., Ambikairajah, E., Lovell, N. H. and Celler, B. G., 2006. Classification of a known sequence of motions and postures from accelerometry data using adapted Gaussian mixture models. Physiological measurement, 27(10), p.935.

67. Marmanis, D., Datcu, M., Esch, T. and Stilla, U., 2016. Deep learning earth observation classification using ImageNet pretrained networks. IEEE Geoscience and Remote Sensing Letters, 13(1), pp.105-109.

68. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C. and Lawrence Zitnick, C., 2015. From captions to visual concepts and back. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1473-1482).

69. Deng, L. and Jaitly, N., 2015. Deep discriminative and generative models for pattern recognition. USENIX–Advanced Computing Systems Association.

70. Giri, E.P., Fanany, M.I. and Arymurthy, A.M., 2016. Combining Generative and Discriminative Neural Networks for Sleep Stages Classification. arXiv preprint arXiv:1610.01741.

71. Hochreiter, S. and Schmidhuber, J., 1997. Long short-term memory. Neural computation, 9(8), pp.1735-1780.

72. Britz, D., 2015. Recurrent Neural Networks Tutorial, Part 1–Introduction to RNNs.

73. Graves, A., 2012. Supervised sequence labelling with recurrent neural networks (Vol. 385). Heidelberg: Springer.

74. Noguchi, W., Iizuka, H. and Yamamoto, M., 2016, May. Proposing Multimodal Integration Model Using LSTM and Autoencoder. In proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS) on 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS) (pp. 355-362). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

75. Wang, J, 2014. A Tutorial on Speaker Verification. Center for Speech and Language Technologies. Tsinghua University. http://cslt.riit.tsinghua.edu.cn/mediawiki/images/c/cb/131104-ivector-microsoft-wj.pdf

76. American Clinical Neurophysiology Society, 2006. Guideline 3: minimum technical standards for EEG recording in suspected cerebral death. J Clin Neurophysiol, 23, pp.97-104.

77. Marco Bellantonio, 2016. Hybrid CNN+LSTM for Face Recognition in Videos. Thesis Presentation. Departamento de Informática, Universidad Técnica Federico Santa María (UTFSM). http://sergioescalera.com/wp-content/uploads/2016/12/Thesis_presentation.pdf

78. Gómez-Herrero, G., De Clercq, W., Anwar, H., Kara, O., Egiazarian, K., Van Huffel, S. and Van Paesschen, W., 2006, June. Automatic removal of ocular artifacts in the EEG without an EOG reference channel. In Signal Processing Symposium, 2006. NORSIG 2006. Proceedings of the 7th Nordic (pp. 130-133). IEEE.

79. Vialatte, F.B., Solé-Casals, J. and Cichocki, A., 2008. EEG windowed statistical wavelet scoring for evaluation and discrimination of muscular artifacts. Physiological Measurement, 29(12), p.1435.

80. Hu, J., Wang, C.S., Wu, M., Du, Y.X., He, Y. and She, J., 2015. Removal of EOG and EMG artifacts from EEG using combination of functional link neural network and adaptive neural fuzzy inference system. Neurocomputing, 151, pp.278-287.

81. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A. and Liu, P., 2013, April. A high-resolution spontaneous 3d dynamic facial expression database. In Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on (pp. 1-6). IEEE.

82. Dhineshkumar, R., Ganesh, A.B. and Sasikala, S., 2016. Speaker identification system using Gaussian mixture model and support vector machines (GMM-SVM) under noisy conditions. Indian Journal of Science and Technology, 9(19).

83. Palazzo, S., Spampinato, C., Kavasidis, I., Giordano, D. and Shah, M., 2017. Generative Adversarial Networks Conditioned by Brain Signals. PDF available on ucf.edu.

84. Spampinato, C., Palazzo, S., Kavasidis, I., Giordano, D., Shah, M. and Souly, N., 2016. Deep Learning Human Mind for Automated Visual Classification. arXiv preprint arXiv:1609.00344.

85. Kessous, L., Castellano, G. and Caridakis, G., 2010. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. Journal on Multimodal User Interfaces, 3(1), pp.33-48.

86. Gupta, P. and Rajput, N., 2007. Two-stream emotion recognition for call center monitoring. In Eighth Annual Conference of the International Speech Communication Association.

87. Fragopanagos, N. and Taylor, J.G., 2005. Emotion recognition in human–computer interaction. Neural Networks, 18(4), pp.389-405.

88. Vogt, T. and André, E., 2005, July. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on (pp. 474-477). IEEE.

89. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W.F. and Weiss, B., 2005, September. A database of german emotional speech. In Interspeech (Vol. 5, pp. 1517-1520).

90. Soleymani, M., Pantic, M. and Pun, T., 2012. Multimodal emotion recognition in response to videos. IEEE transactions on affective computing, 3(2), pp.211-223.

91. Ren, J.S., Hu, Y., Tai, Y.W., Wang, C., Xu, L., Sun, W. and Yan, Q., 2016, February. Look, Listen and Learn-A Multimodal LSTM for Speaker Identification. In AAAI (pp. 3581-3587).

92. Zhang, Z., Girard, J.M., Wu, Y., Zhang, X., Liu, P., Ciftci, U., Canavan, S., Reale, M., Horowitz, A., Yang, H. and Cohn, J.F., 2016. Multimodal spontaneous emotion corpus for human behavior analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3438-3446).

93. Shan, C., Gong, S. and McOwan, P.W., 2009. Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing, 27(6), pp.803-816.

94. Yacoub, S.M., Simske, S.J., Lin, X. and Burns, J., 2003, September. Recognition of emotions in interactive voice response systems. In INTERSPEECH.

95. Mollahosseini, A., Hasani, B. and Mahoor, M.H., 2017. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. arXiv preprint arXiv:1708.03985.

96. Vogt, T., André, E. and Bee, N., 2008. EmoVoice—A framework for online recognition of emotions from voice. Perception in multimodal dialogue systems, pp.188-199.

97. Gunes, H. and Piccardi, M., 2007. Bi-modal emotion recognition from expressive face and body gestures. Journal of Network and Computer Applications, 30(4), pp.1334-1345.

98. Crane, E. and Gross, M., 2007. Motion capture and emotion: Affect detection in whole body movement. Affective computing and intelligent interaction, pp.95-101.

99. Bernhardt, D., 2010. Emotion inference from human body motion (Doctoral dissertation, University of Cambridge).

100. Piana, S., Stagliano, A., Odone, F., Verri, A. and Camurri, A., 2014. Real-time automatic emotion recognition from body gestures. arXiv preprint arXiv:1402.5047.

101. Gouizi, K., Bereksi Reguig, F. and Maaoui, C., 2011. Emotion recognition from physiological signals. Journal of medical engineering & technology, 35(6-7), pp.300-307.

102. Uma, I., 2014. physiological signals based human emotion recognition a review. American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS), 6(1).

103. Bänziger, T., Grandjean, D. and Scherer, K.R., 2009. Emotion recognition from expressions in face, voice, and body: the Multimodal Emotion Recognition Test (MERT). Emotion, 9(5), p.691.

104. Liu, W., Zheng, W.L. and Lu, B.L., 2016. Multimodal emotion recognition using multimodal deep learning. arXiv preprint arXiv:1602.08225.

105. Parlak, C. and Diri, B., 2013, April. Emotion recognition from the human voice. In Signal Processing and Communications Applications Conference (SIU), 2013 21st (pp. 1-4). IEEE.

106. Nakasone, A., Prendinger, H. and Ishizuka, M., 2005, September. Emotion recognition from electromyography and skin conductance. In Proc. of the 5th International Workshop on Biosignal Interpretation (pp. 219-222).

107. Yang, S. and Yang, G., 2011. Emotion Recognition of EMG Based on Improved LM BP Neural Network and SVM. JSW, 6(8), pp.1529-1536.

108. Jerritta, S., Murugappan, M., Wan, K. and Yaacob, S., 2014. Emotion recognition from facial EMG signals using higher order statistics and principal component analysis. Journal of the Chinese Institute of Engineers, 37(3), pp.385-394.

109. Kothe, C.A., Makeig, S. and Onton, J.A., 2013, September. Emotion recognition from EEG during self-paced emotional imagery. In Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on (pp. 855-858). IEEE.

110. Blaiech, H., Neji, M., Wali, A. and Alimi, A.M., 2013, December. Emotion recognition by analysis of EEG signals. In Hybrid Intelligent Systems (HIS), 2013 13th International Conference on (pp. 312-318). IEEE.

111. Lahane, P. and Sangaiah, A.K., 2015. An approach to EEG based emotion recognition and classification using kernel density estimation. Procedia Computer Science, 48, pp.574-581.