Unsupervised Affective State Learning from Speech

by

Gökhan Kuşçu

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

Master of Science

in

Electrical and Electronics Engineering

KOÇ ÜNİVERSİTESİ

May 9, 2024

Unsupervised Affective State Learning from Speech

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Gökhan Kuşçu

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

Committee Members:

_____

Prof. Engin Erzin (Advisor)

_____

Prof. Yücel Yemez

_____

Prof. Çiğdem Eroğlu Erdem

Date: _____

To those who encourage what is good and forbid what is evil.

ABSTRACT

Unsupervised Affective State Learning from Speech
Gökhan Kuşçu
Master of Science in Electrical and Electronics Engineering
May 9, 2024

The conventional paradigm for estimating continuous emotional states from speech, investigated as a regression problem over time, has been widely acknowledged. This thesis introduces a novel methodology that transposes this challenge into the classification domain by learning clusters of affect contours of uniform lengths. Our approach involves a novel joint clustering and classification scheme, wherein each iteration involves clustering affect contours into independent classes. We seek to classify these classes, identifying distinct clusters with observed intra-class sample similarities. The classification structure integrates an audio feature extractor based on a Wav2Vec 2.0 model, followed by a convolutional neural network (CNN). Concurrently, the clustering component processes a segment of the affect contour, employing a convolutional network for dimensionality reduction and subsequent application of k-means clustering. The classification network predicts these generated clusters. The cumulative loss is then propagated to neural networks for weight updates. Empirical findings reveal that the obtained clusters exhibit distinctive and insightful characteristics. Simultaneously, incorporating a regression head into the trained classification network yields competitive audio-only performance on the RECOLA and USC CreativeIT datasets regarding continuous emotion recognition (CER). The results for CER are compared against baselines and existing literature, illustrating the efficacy of our approach. Our results demonstrate that while achieving competitive continuous emotion recognition performance, our approach, fundamentally a classification framework, converts the nature of the well-studied continuous regression problem.

ÖZETÇE

Konuşmadan Gözetimsiz Duygusal Durum Öğrenme
Gökhan Kuşçu
Elektrik ve Elektronik Mühendisliği, Yüksek Lisans
9 Mayıs 2024

Konuşmadan sürekli duygu durumlarını tahmin etmeye yönelik geleneksel paradigma, zaman içinde bir regresyon problemi olarak yorumlanmış ve yaygın olarak kabul görmüştür. Bu tez, tek tip uzunluklara sahip duygu konturları kümeleri oluşturarak bu zorluğu sınıflandırma alanına aktaran yeni bir metodoloji sunmaktadır. Bu yaklaşım, her yinelemede duygulanım konturlarının bağımsız sınıflar halinde kümelenmesini içeren yeni bir ortak kümeleme ve sınıflandırma şeması içermektedir. Gözlemlenen sınıf içi örnek benzerlikleri ile farklı kümeleri tanımlayarak bu sınıfları sınıflandırmaya çalışılmaktadır. Sınıflandırma yapısı, Wav2Vec 2.0 modeline dayalı bir ses özelliği çıkarıcıyı ve ardından bir evrişimli sinir ağını (CNN) içermektedir. Eş zamanlı olarak, kümeleme bileşeni, boyutsallığı azaltmak ve ardından k-means kümelemesini uygulamak için bir evrişimli ağ kullanarak etki konturunun eşit uzunluktaki bir birimini işler. Kümeleme bileşeni tarafından üretilen kümeler sınıflandırma ağı tarafından tahmin edilir. Kümülatif kayıp daha sonra ağırlık güncellemeleri için sinir ağına yayılır. Ampirik bulgular, elde edilen kümelerin birbirinden farklı ayırt edici özellikler sergilediğini ortaya koymaktadır. Eş zamanlı olarak, eğitilmiş sınıflandırma ağına bir regresyon başlığının dahil edilmesi, RECOLA ve USC CreativeIT veri kümelerinde sadece ses kullanıldığında literatürdeki sonuçları yakalayan bir performans sağlar. Bu sonuçlar, regresyon baz performansları ve mevcut literatür ile karşılaştırılarak yaklaşımımızın etkinliği gösterilmiştir. Temelde bir sınıflandırma çerçevesi olan yaklaşımımız rekabetçi sürekli duygu tanıma performansına ulaşırken iyi çalışılmış sürekli regresyon problemi doğasını dönüştürmektedir.

ACKNOWLEDGMENTS

I express sincere gratitude to Prof. Engin Erzin for his unwavering support and invaluable guidance in completing this master's thesis. His wisdom and mentorship shaped the research, and I deeply appreciate his encouragement during challenging moments. This thesis represents both an academic accomplishment and a testament to his impactful mentorship. I am thankful for his attention, sacrifices, and extensive knowledge.

I extend my heartfelt appreciation to the members of the thesis defence committee, Prof. Yücel Yemez and Prof. Çiğdem Eroğlu Erdem, for their valuable insights, constructive feedback, and dedication during the evaluation of this work. I am grateful for their time, consideration, and the collaborative spirit they brought to the defence process.

I am deeply grateful to my loving family for their constant support and encouragement in my academic pursuits.

In my master's journey, I am profoundly thankful to my wife, Afife Beyza, for her steadfast support. Her strength and encouragement have been indispensable contributors to my success.

This thesis is partially supported by The Scientific and Technological Research Council of Türkiye (TÜBİTAK) BIDEB 2210-A Scholarship program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

SER	Speech Emotion Recognition
CNN	Convolutional Neural Network
GRU	Gated Recurrent Unit
SVM	Support Vector Machine
GMM	Gaussian Mixture Model
RBF	Radial Basis Functions
kNN	k-Nearest Neighbors
GAN	Generative Adversarial Network
LMT	Logistic Model Tree
HMM	Hidden Markov Model
LSTM	Long Short-Term Memory
CCC	Concordance Correlation Coefficient
RMSE	Root Mean Squared Error
MFB	Mel Filter Bank
AVD	Arousal-Valence-Dominance
CER	Continuous Emotion Recognition
MFCC	Mel-Frequency Cepstrum Coefficients
BERT	Bidirectional Encoder Representations from Transformers
RECOLA	Remote Collaborative and Affective Interactions

Chapter 1

INTRODUCTION

1.1 Introduction to Speech Emotion Recognition

Human emotions constitute a fundamental aspect of communication, providing depth and meaning to our interactions. Gestures, voice tones, and facial expressions naturally convey these emotions, easily understood by humans. However, when machines attempt to understand elements like spoken content, written text, or audiovisuals, communicating emotional knowledge becomes a formidable challenge. Machines and intelligent systems, lacking an inherent understanding of emotions, face difficulty interpreting the nuanced emotional content embedded in these forms of communication.

As individuals extensively engage in verbal communication as a highly convenient means of interaction, replicating this human behaviour in communicating with machines has become a significant advancement in recent years. Beyond merely conveying information in a manner understandable to machines, there is a growing emphasis on incorporating the natural human inclination to speak and seek understanding in terms of conveyed content and the accompanying emotions. This becomes increasingly vital in facilitating human-machine interactions where emotional expression plays a key role alongside the conveyed message.

Since human-computer interaction advances and becomes increasingly integral to daily life, the research community focuses on speech emotion recognition (SER), promising to enhance natural communication between people and computer systems. The applications of speech emotion recognition span various domains, including driver assistance, healthcare, entertainment, chatbots, and more. The insights gained from this technology facilitate more natural interactions and offer valuable

information on interpreting human behaviour by intelligent systems. With the evolution of innovative systems, individuals now engage in more genuine communication through various mediums, including visual and auditory channels. Exploring multimodal systems, where machines seamlessly integrate multiple modalities, is a prominent and widely pursued research direction in this context.

Speech emotion recognition has been investigated through various research. Both traditional machine learning methods and deep neural networks have been employed to tackle the classification challenge in speech emotion recognition, with a notable use of the latter in recent years. In studies such as [Gao et al., 2017], [Dahake et al., 2016], and [Milton et al., 2013], authors applied different linear kernels of Support Vector Machine (SVM), where they initially extracted Mel-Frequency Cepstrum Coefficients (MFCC) from spoken utterances and normalized the obtained signals before inputting them into SVMs. SVMs have emerged as the most frequently cited conventional machine learning algorithm and are extensively utilized in this context.

Beyond SVMs, other conventional methods are also explored. For instance, in [Akash et al., 2016], the authors utilized Multilayer Perceptron (MLP), Bayes networks, and SVMs, comparing their performance on original data and its spectrograms. [Kishore and Satish, 2013] examined the impact of using MFCC features and wavelets with Gaussian Mixture Model (GMM), while [Neiberg et al., 2006] employed GMM to classify emotions using MFCC features calculated at different frequency levels. In [Kerkeni et al., 2019], optimal feature selection efforts involved a combination of SVMs and Recurrent Neural Networks (RNN) for emotion classification, with feature sets generated by various methods such as wavelets and MFCC, also being compared in [Wang et al., 2020] using SVMs with radial kernels. [Rieger et al., 2014] and [Abdel-Hamid, 2020] utilized the k-Nearest Neighbour (kNN) algorithm to address the classification task, investigating the use of spectral features and MFCCs. Linear, polynomial, and Radial Basis Function (RBF) kernels are evaluated in [Chavhan et al., 2010], where the effectiveness of both kernels is examined post-extraction of MFCC features.

Additionally, logistic model tree (LMT) classifiers were applied to unnormalized MFCC features in [Zamil et al., 2019]. Conventional classifiers, including kNN, SVM,

and MLP, were compared to a model formed by maximum likelihood estimation (MLE) in [Zhao et al., 2013]. Random Forest (RF) and MLP performance were compared on MFCC features in [Iliou and Anagnostopoulos, 2009]. Dimensionality reduction methods were employed, with [Hou et al., 2020] using non-negative matrix factorization (NMF) and [Arias et al., 2014] using Principal Component Analysis (PCA) before feeding data to the model. In [Mao et al., 2009], a joint model of the Hidden Markov Model (HMM) and neural networks addressed the classification task. [Hou et al., 2020] presented a recognition system using Gradient Boosting (GB) on MFCC and spectral features.

Neural network-based approaches were explored as well. LSTM networks were applied for emotion classification in [Caponetti et al., 2011], and a model based on a 1-dimensional Convolutional Neural Network (CNN) was proposed on MFCC features in [Issa et al., 2020]. [Zhang et al., 2021] employed a combination of CNNs with different dimensionalities, processing audio recordings with 1-dimensional CNNs, MFCC features with 2-dimensional CNNs, and utilizing 3-dimensional CNNs for temporal and spatial properties of spoken content. Both neural networks and conventional methods were investigated on datasets and corpora, as seen in [Shahin et al., 2022], where CNN and LSTM architectures were compared to traditional machine learning models. [Andayani et al., 2022] proposed a model based on transformers and LSTMs, while in [Praseetha and Vadivel, 2018], a combination of Gated Recurrent Units (GRU) formed a recurrent network on data after extracting features with MFCCs.

Transfer learning methods were also employed. In [Huang and Bao, 2019], a pretrained AlexNet [Krizhevsky et al., 2012] is utilized to create a neural network for processing MFCC features, while [Wang et al., 2022] combined VGG [Simonyan and Zisserman, 2015] architecture with LSTM networks. Similarly, a model using ResNet [He et al., 2015] architecture recognized emotion in audio content. In [Jannat et al., 2018], Inception-V3 [Szegedy et al., 2015] was employed for a multimodal emotion recognition task. Finally, [Bakhshi et al., 2022] introduced a data augmentation step using a Generative Adversarial Network (GAN) on raw audio before feeding it into a CNN architecture for audio recording classification.

The following sections firstly provide a brief overview of the commonly used databases employed in speech emotion recognition research. Afterwards, we present details about the emotion model established by [Thayer, 1989], which translates discrete emotion states into dimensions of energy and stress. The dataset utilized in this study is generated per this model.

1.1.1 Common Databases

The EMODB [Burkhardt et al., 2005] is a German database comprising audio recordings featuring ten actors (5 male and 5 female). These actors produce 10 German utterances, encompassing 5 short and 5 longer sentences, expressing 7 emotions (anger, neutral, anger, boredom, happiness, sadness, disgust). The database prioritizes sentences for a more natural form, allowing actors to speak from memory. eNTERFACE [Martin et al., 2006] is an audiovisual emotion database that includes six emotions expressed by 42 subjects from 14 nationalities in reaction to six successive short stories. With a focus on English, the database has been utilized to test both conventional machine learning and deep learning models. Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [Busso et al., 2008] recorded from ten actors in dyadic sessions features markers on the face, head, and hands during scripted and improvised scenarios. Initially targeting five emotions, it expanded to 10 categories. INTER1SP comprises recordings of one male and one female Spanish professional actor, totalling 184 utterances with emotions like anger, disgust, fear, happiness, sadness, surprise, and neutral. MSP-PODCAST [Lotfian and Busso, 2019] provides an extensive emotional database using existing spontaneous recordings from audio-sharing websites. It delivers over 18,000 natural emotional sentences from multiple speakers, annotated with emotional attributes. Toronto Emotional Speech Set (TESS) [Pichora-Fuller and Dupuis, 2020] includes about 2,800 samples featuring seven emotions. It is frequently used for testing deep learning models and is applied in conventional machine learning methods. RECOLA [Ringeval et al., 2013] involves 9.5 hours of audio, visual, and physiological recordings of dyadic interactions. French-speaking participants collaborated on a task, expressing affective and social

Table 1.1: Common databases and their properties (AVM: Audio-Visual-Motion, AVD: Arousal-Valence-Dominance, Act: Acted, Spon: Spontaneous, D: Discrete, C: Continuous)

Database	**Lang.**	**Modality**	**Type**	**Annotation**	**Duration**
EMODB	German	Audio	Act	D: 7-class	800 everyday sentences (4.15 hour-long)
eNTERFACE	English	Audio-Visual	Act	D: 6-class	1116 utterances (6 successive short stories)
IEMOCAP	English	AVM	Act	D: 10-class	12 hours of audiovisual recording
INTER1SP	Spanish	Audio	Act	D: 7-class	184 utterances (4 hour-long)
MSP-PODCAST	English	Audio	Spon	C: AVD	17500 sentences (27 hour-long)
TESS	English	Audio	Act	D: 7-class	2800 sentences
CreativeIT	English	AVM	Spon	D & C	2-10 minutes of dyadic interactions
RECOLA	French	Audio-Visual	Spon	C: AVD	27 utterances (3.8 hour dyadic interactions)

behaviours, which participants and annotators annotated. The database contains both continuous and multimodal data. CreativeIT [Metallinou et al., 2010] is based on the active analysis improvisation technique and includes 50 dyadic theatrical improvisations performed by 16 actors, providing full-body motion capture data and audio-video data for studying theatrical performance and human communication.

1.1.2 Arousal-Valence Emotional State Model

In speech emotion recognition, the formation of target values, also known as ground truth values, significantly influences the process. Traditional emotions such as happiness, anger, and sadness rely on human perceptions for label creation, which is also characterized in [Ekman et al., 1987] to distinguish human emotions into six archetypal emotional states, namely happiness, surprise, sadness, fear, disgust, anger, research efforts aim to determine and illustrate various emotions based on shared information in a lower-dimensional space. In [Thayer, 1989], the author introduced a straightforward and effective two-dimensional emotion model that categorizes emotional responses into stress and energy. This model, as depicted in Figure 1.1, designates the stress dimension as valence and the energy dimension as arousal, en-

suring the independence of information carried by these two axes. From a human perception standpoint, this implies that arousal values are not dependent on valence values and vice versa. Every human emotion can be numerically dissociated into distinct arousal-valence pairs, allowing for their representation on a 2-dimensional diagram based on the specific location corresponding to each emotion.

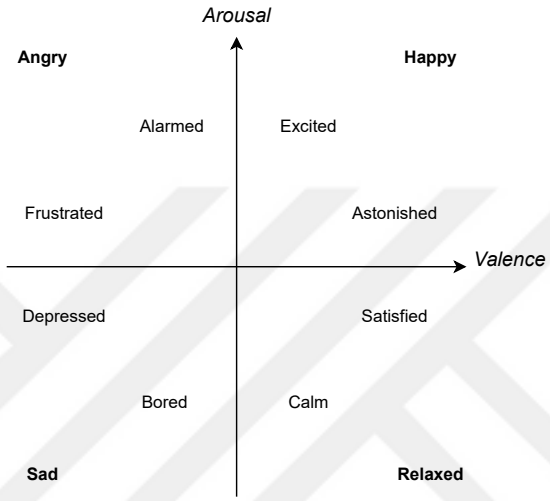


Figure 1.1: *2D* arousal-valence diagram with emotions

1.2 Literature Review

This section will compile recent literature focused on audio-only emotion recognition, specifically targeting arousal and valence datasets.

In the paper [Khorram et al., 2017a], the authors explored two architectures that provide different methods for capturing long-term temporal dependencies in a given sequence of acoustic features. Dilated convolutions incorporate long-term temporal information without changing the input signal's length by filters with varying dilation factors. Down/upsampling networks are used for long-term dependencies by applying a series of convolutions and max poolings to downsample the signal to obtain a global picture of the features. The downsampled signal is then utilized to reconstruct the output with a length equal to the uncompressed input. In the paper [AlBadawy and Kim, 2018], a joint modelling method merges discrete and contin-

uous emotion representations. Some studies reported enhancements in continuous emotion prediction performance by quantizing noisy continuous labels and employing these quantized, discrete labels for training classification models. Quantization approaches can introduce errors when converting continuous labels to discrete ones. To address this challenge and determine the optimal balance between discretized and continuous emotion representations, two joint modelling methods are introduced: ensemble and end-to-end. The ensemble method combines outputs from the classification and regression models at the prediction phase. In contrast, the end-to-end model is trained to optimize classification, regression, and the combined tasks simultaneously in an end-to-end manner. These models are based on D-BLSTM architectures. [Khorram et al., 2021] employs an approach to aligning acoustic features and continuous emotion labels utilizing the delayed sinc layer. It is a significant concern when aligning speech and emotion annotations due to variations in reaction delays among annotators based on affective behaviours. Multiple delays occur since layers are combined into a network architecture to address this. The network categorizes features into fuzzy clusters, each acoustic feature belonging to more than one cluster, and then learns different delays for different acoustic clusters. The paper focused on utilizing the speech modality for predicting continuous emotion annotations as the proposed multi-delay sinc network is reported to have a modelling technique for other input modalities. [Fatima and Erzin, 2021] introduce the concept of dyadic context and propose CNN to enhance CER in dyadic settings. They first present a CNN model for single-subject CER using speech and body motion data, followed by a two-stage regression framework for dyadic CER. Additionally, they propose a new Convolutional LSTM model for dyadic continuous emotion recognition. They report performance enhancements on the USC CreativeIT database, with results showing favourable comparisons to state-of-the-art approaches. The authors [Lin et al., 2021] introduced the DeepEmoCluster framework, a semi-supervised approach that employs latent representations for Speech Emotion Recognition (SER) using labelled and unlabeled datasets. Through a supervised emotional regression task, the DeepEmoCluster approach constructs a latent feature space based on clusters affected by emotional content. Including a maximum latent cluster constraint in the joint

training process is reported to allow the approach to attain optimal prediction scores for arousal and competitive performance for valence in a fully supervised training setup. Another study [Lin and Busso, 2024] introduces the temporal DeepEmoCluster framework, which utilizes temporal constraints in the utilization of clusters as an additional task for semi-supervised learning (SSL) tasks. Two sentence-level temporal modelling approaches, the temporal network and the triplet loss function, are employed to impose temporal constraints. The temporal network method involves an additional network module to encode temporal information across data chunks in a sentence. The triplet loss function utilizes the temporal relationship between data chunks to learn hidden representations with additional temporal constraints. A sentence-level uniform sampling strategy is proposed to maintain the original temporal orders of data chunks in a sentence. Experimental results on the audio-only corpus report that the temporal models outperform the vanilla DeepEmoCluster and other existing reconstruction-based semi-supervised frameworks in speech emotion recognition (SER) tasks across emotion attributes like arousal and valence.

1.3 Contribution of thesis

This paper introduces a novel approach to speech emotion recognition, focusing on identifying temporal affect contour clusters that exhibit high predictability from speech representations, diverging from traditional regression and classification tasks. To address this, we propose a novel joint clustering and classification network to uncover novel affect contour clusters in speech data with increased precision. The framework comprises key elements: the affect network, responsible for extracting latent representations of affect contours and associating them with affect clusters; the speech network, tasked with extracting latent representations of speech signals and linking them to affect clusters; and the clustering block, which executes k-means clustering updates on the latent representations of affect contours. It's worth noting that while similarities exist between the DeepEmoCluster and our proposed frameworks, our proposed framework distinguishes itself with particular features. Our framework generates pseudo-labelled clusters from the latent representations

of affect contours, bypassing direct reliance on speech, and employs a network that predicts affect contour cluster labels from speech data, reframing it as a classification challenge. Moreover, specific discretization techniques have been utilized for converting regression into classification, employing equal-length segments for both speech and affect data represents a fresh approach. Experimental results validate the effectiveness of this approach in classification and demonstrate its capability to improve regression performance by leveraging insights obtained during classification tasks.

Chapter 2

METHODOLOGY

This chapter describes our proposed joint neural network, comprising distinct classification and clustering sub-networks, and an in-depth explanation of their respective characteristics. Subsequently, an account of the regression head employed for assessing the effectiveness and performance of the proposed methodology relative to baseline networks in the existing literature is presented.

2.1 Joint Classification and Clustering Network

Emotion recognition is a prominent research domain, marked by the proliferation of diverse approaches to address this complex challenge. A pivotal focus within this domain is continuous emotion recognition, conceptualized as a regression problem. In this research, we aim to identify patterns in the temporal expression of emotions, known as affect contours, that can be reliably inferred from speech features. We present a new method that integrates clustering and classification networks to achieve this goal. Our approach reveals affect contour clusters in speech data, improving predictive accuracy.

The proposed joint clustering and classification network comprises two primary components: an affect network and a speech network. The speech network (SpeechNet) inputs speech signals and predicts labels for affect contour clusters. The affect network (AffectNet) operates on affect contours and serves two functions. Firstly, it extracts a lower-dimensional latent affect representation and updates the centroids defining the affect contour cluster labels using k-means. Secondly, it predicts affect contour cluster labels based on the latent affect representations. The goal is to categorize affect contours into separate classes and predict them using a classification network with speech data as input.

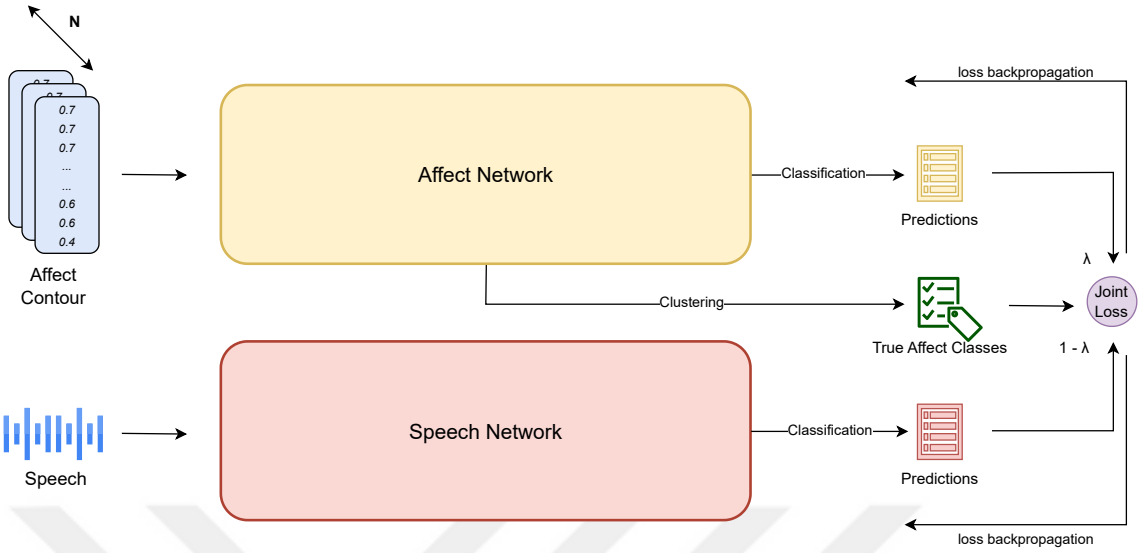


Figure 2.1: Higher level block diagram of the Joint Classification and Clustering Network (JCCN) consisting of affect and speech networks

The initial network, denoted as SpeechNet, is fed with speech data and is structured as a fusion of an audio feature extraction network and a convolutional neural network (CNN). The audio extraction component transforms speech data into corresponding audio representations, while the CNN component produces classification results for the entire network. SpeechNet thus serves as a classifier mapping speech to affect clusters. The second network, referred to as AffectNet, accepts affect contours as input and is designed as a CNN classifier akin to its counterpart. AffectNet generates affect cluster predictions at its output layer and, through intermediate layers, captures latent representations of affect contours. These representations are clustered to form ground truth affect clusters, which both the speech classifier and affect classifier networks predict.

During training, losses from both networks are aggregated and backpropagated for weight updates, except for the frozen audio feature extraction part. The loss function for both networks is defined as multi-class cross-entropy loss.

$$L = \alpha L_{speech} + (1 - \alpha) L_{affect} \tag{2.1}$$

Algorithm 1: Proposed algorithm for Joint Classification and Clustering Network (JCCN)

Input: Speech data, affect contour (RECOLA)
Output: Trained SpeechNet, AffectNet networks

```
// Segment data, 2-second long speech segments with 1-second intercepts
// and 50-unit long affect contour samples with 25-unit long intercepts
speechSegs, affectSegs ← SegmentSpeech(speech), SegmentAffect(affectContour);

// Initialize models
AffectNet, SpeechNet ← InitAffectNet(), InitSpeechNet();

// Epoch Loop
for epoch ← 1 to totalEpochs do
    // AffectNet to produce representations with frozen layers
    _, latentReps ← AffectNetForward(affectSegs);

    // Clustering latent representations to generate affect classes
    clusters ← KMeansClustering(latentReps);

    // Training Loop
    for batch ← 1 to totalBatches do
        // AffectNet forward
        affectPreds, _ ← AffectNetForward(affectSegs[batch]);

        // SpeechNet forward
        speechPreds ← SpeechNetForward(speechSegs[batch]);

        // Calculate loss and backpropagate
        totalLoss ← CalculateTotalLoss(affectPreds, speechPreds, clusters);
        Backpropagate(AffectNet, SpeechNet, totalLoss);
// Trained models
return AffectNet, SpeechNet;
```

The loss function for the combined network is defined in Equation 2.1, where α represents a weighting coefficient, and L_{speech} and L_{affect} denote the respective cross-entropy losses of the SpeechNet and AffectNet. We conduct a grid search to determine the ideal value for α to achieve the most favourable classification outcomes.

Throughout the training process, both the network parameters and the k-means clusters of the latent affect representations are iteratively updated. The latent affect representations go through k-means clustering once per epoch to ensure consistency across batches. These clusters are then utilized for speech segment processing, which is subsequently inputted into the joint clustering and classification network in batches. The network achieves smoother convergence by separating the clustering step from the training loop and initializing subsequent epochs with clustering cen-

troids. Another aspect of the clustering process is that since the model weights of the affected network are updated for every training epoch, the latent representation is modified between these epochs. To obtain consistent cluster centroids between epochs and better convergence during the training, we feed the clustering part with the previous cluster centroids obtained in the last loop. This addition to the clustering process ensures a better convergence and also allows for keeping track of the generation of the clustering centroids.

2.1.1 Speech Network

SpeechNet operates in two phases to predict affect contour cluster labels from the speech input. The first phase utilizes a pre-trained acoustic feature extractor. In the second phase, it maps high-dimensional acoustic feature embeddings to low-dimensional acoustic latent representations and subsequently to the affect contour cluster labels.

The speech network consists of two sub-networks. The first sub-network is a pre-trained audio feature extractor, while the second is a multilayer convolutional neural network that predicts classes. The audio feature extraction phase prepares the audio data by extracting pertinent information for subsequent processing. Various techniques have been employed for audio feature extraction, such as applying mel filter bank coefficients as detailed in the works [Khorram et al., 2021] and [Khorram et al., 2017b]. Both methods extract 40-dimensional log Mel Filter Bank (MFB) features using a window length of 25ms, a benchmark parameter tuned to obtain audio representations. Conversely, another widely adopted method involves the utilization of Mel-frequency Cepstrum Coefficients (MFCC), as utilized in the paper [Zhang et al., 2018]. Prior research [Busso et al., 2007] has demonstrated the superior efficacy of MFB features compared to traditional MFCCs for emotion prediction.

For audio representation extraction in this research, a departure from established literature prompted the adoption of Wav2Vec [Baevski et al., 2020]. Following the transformer [Vaswani et al., 2017] paradigm akin to BERT [Devlin et al., 2018], the model undergoes training by predicting speech units for selectively masked seg-

ments within the audio. Initially, the raw waveform of the speech audio undergoes processing through a multilayer convolutional neural network to yield latent audio representations. These representations are subsequently input to both a quantizer and a transformer. A subset of the audio representations is subjected to masking before integration into the transformer. The transformer assimilates information from the complete audio sequence. Ultimately, the transformer's output is utilized to tackle a contrastive task, requiring the model to discern the correct quantized speech units for the masked positions. The contextual insights provided by this model underscore its preference for processing speech data.

Table 2.1: Network architecture of the SpeechNet

Layer	**Type**	**Depth**	**Kernel**	**Output**
1	Wav2Vec	-	-	99x1024x1
2	CONV+ReLU	32	7x7	93x1018x32
3	MaxPooling	32	2x2	46x509x32
4	CONV+ReLU	64	7x7	40x503x64
5	MaxPooling	64	2x2	20x251x128
6	CONV+ReLU	128	5x5	16x247x128
7	CONV+ReLU	256	5x5	12x243x256
8	MaxPooling	256	2x2	6x121x512
9	CONV+ReLU	512	3x3	4x119x512
10	CONV+ReLU	512	3x3	2x59x512
11	MaxPooling	512	2x2	1x29x512
12	FC+LeakyReLU	-	-	1024
13	FC+LeakyReLU	-	-	512
14	FC+LeakyReLU	-	-	128
15	FC+LeakyReLU	-	-	64
16	FC+Softmax	-	-	4

In this study, we analyze a temporal window of two seconds to derive audio features. The raw audio embeddings are obtained from the pre-trained Wav2Vec 2.0 *Large* [Baevski et al., 2020] model, specifically extracted from the intermediate layer of block 15. The resultant embeddings from this configuration exhibit a shape

denoted as M and D, where M represents the number of acoustic features for the designated two-second audio segment, amounting to 99, and D signifies the dimensionality of the acoustic feature, measuring 1024. The primary rationale for selecting the 15th block is its demonstrated performance, leading to a lower phoneme error rate as evidenced in [Baevski et al., 2021].

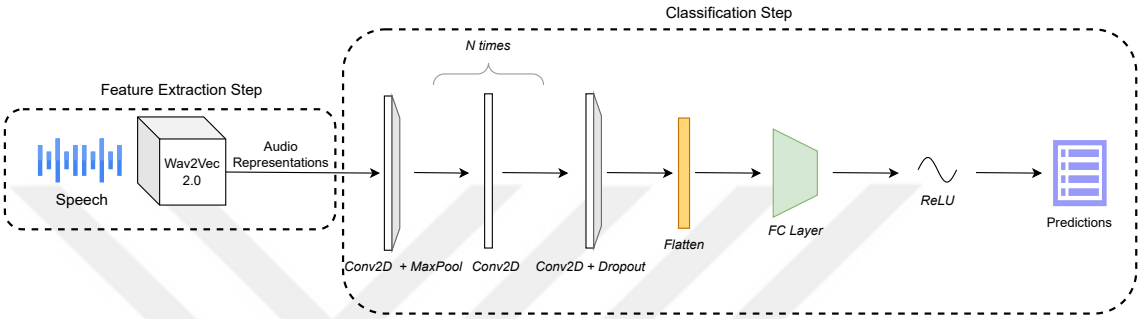


Figure 2.2: An overview of the SpeechNet network to generate affect to predict affect cluster classes through AffectNet processing

As mentioned in the preceding section, the speech network comprises two subnetworks: the initial one is the audio feature extractor, and the subsequent one is a multilayer convolutional neural classifier network. The convolutional network processes these representations to yield lower-dimensional features, producing a denser representation when supplied with the representations acquired in the initial phase. Subsequently, the network generates predictions for affect cluster classes by utilizing fully connected layers in the latter stages.

2.1.2 Affect Clustering Network

Similar to the construction of the speech network, our affect clustering network is designed as a convolutional neural network with fully connected layers responsible for predicting affect classes. This network operates with dual objectives. The primary objective involves acting as an autoencoder [Hinton et al., 2011] without the decoder component. The absence of the decoder is compensated by utilizing the classification part to compute the loss and update the weights. This design compels the

convolutional network to generate features that enhance classification performance. The first target is achieved by extracting a lower-dimensional representation of the affect contour and mapping the 50-unit long affect contour into informative latent representations. These latent representations are subsequently clustered into distinct affect classes through k-means clustering [Jin and Han, 2010]. Simultaneously, as a secondary objective, the entire network functions as a classifier, predicting the generated clusters. The loss function, comprising cross-entropy loss for multi-class classification, is aggregated with the speech network loss. Subsequently, the weights of both networks are updated accordingly. Given the task definition of the affect network, the weight convergence occurs at updated points that enhance the model's ability to produce superior latent representations of the affect contour, facilitating improved clustering in subsequent training iterations.

As depicted in the Figure illustrating the affect network, the affect contour undergoes processing within a network constructed through the concatenation of convolutional networks and the incorporation of dropout layers. Notably, the kernel sizes of the convolutional layers in the earlier segments gradually diminish towards the network's conclusion. This deliberate reduction is implemented to ensure a broader receptive field during the initial stages, facilitating the extraction of global features while simultaneously promoting the detailed extraction of local features.

Table 2.2: Network architecture of the AffectNet

Layer	**Type**	**Depth**	**Kernel**	**Output**
1	CONV+ReLU	32	3	32
2	CONV+ReLU	16	3	16
3	CONV+ReLU	8	3	8
4	FC	-	-	4

Regarding the latent representations extracted for the clustering component, we investigated the impact of selecting intermediate layers and identified a trade-off. Deeper layers contain more informative features about the affect network; however,

the lower-dimensional space proves less effective when applying the k-means algorithm. In seeking an optimal balance between these considerations, we identified that a vector dimension of $D = 8$ performs well for both objectives. Consequently, this dimension is chosen for experimental evaluations of the results. To ensure uniform clustering outcomes across batches, the clusters generated in the prior step serve as an initial reference for subsequent clustering iterations. This process involves applying clustering to the latent representations within the affect network.

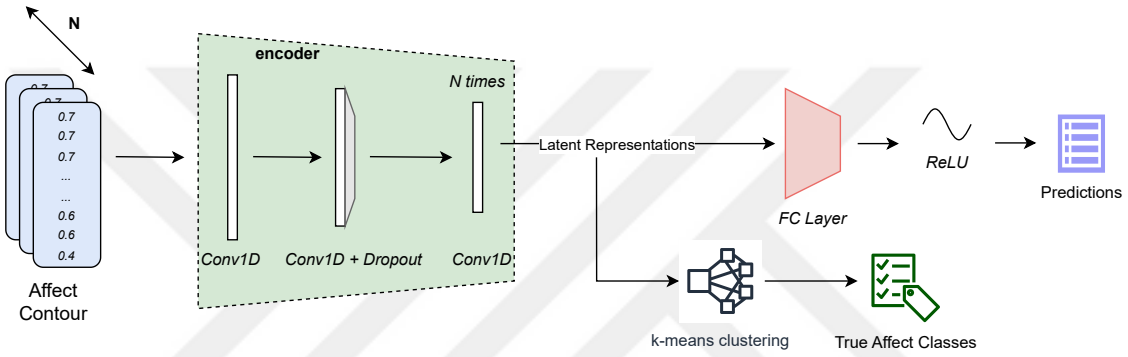


Figure 2.3: Overview of AffectNet network where latent representations are taken from intermediate layer and predictions are generated in output layer

The affect clustering network is initially frozen to ensure the generation of consistent representations for the affect contour. Furthermore, to have a stable clustering across different batches of data, before the start of the training, all data is clustered by applying k-means clustering. After obtaining the clusters for the respective speech segments, batches are created and fed into the network. Placing the clustering step outside the training loop of data batches provides a smoother network convergence, and adaptation is achieved between consecutive batches.

2.2 Continuous Emotion Recognition

Our motivation for undertaking this study is to explore a fresh approach to a regression problem that has undergone extensive examination. As mentioned earlier, the domain of continuous emotion recognition has been thoroughly researched, with

existing literature primarily concentrating on evaluating proposed models' performance in regression. Notable competitions, such as AVEC, notably its AVEC-2016 [Valstar et al., 2016] and AVEC-2018 [Ringeval et al., 2018] editions, have established baseline results and comparison metrics, prompting participants to showcase their outcomes based on these benchmarks. Standard metrics like Root Mean Square Error (RMSE) and Concordance Correlation Coefficient (CCC) are commonly employed to communicate findings and facilitate comparisons with existing literature.

Given our preference toward treating emotion recognition as a classification problem, we encounter challenges in quantitatively assessing the proposed model's contribution beyond presenting classification outcomes. Accordingly, an additional step is integrated into our unified neural network.

This augmentation involves taking the speech network in its entirety, encompassing audio feature extraction and convolutional network components, and fixing the weights post-joint training. Subsequently, we isolate the network to the fully connected layer, characterized by a dimensionality of $D = 128$, determined through experimentation. This subset constitutes a pre-trained segment of the speech network, trained explicitly for affect cluster classification. A regression head is then appended to this segment, predicting the central elements of previously established affect contours from the classification task. The regression head, containing concatenated convolutional layers akin to those in affect and speech networks, is tailored to yield a singular prediction rather than functioning as a classifier. Loss calculation utilizes the mean squared loss (MSE) during network training, with ground truth values comprising the central elements of the affect segment.

This modification reinstates the task definition to a conventional regression paradigm, wherein the network predicts the actual affect contour instead of the cluster classification task, where created classes do not inherently represent ground truth values as observed in previous studies. Consequently, the redefined task aligns with existing literature on regression tasks, enabling comparisons through metrics presented at the start of this section, wherein predicted values are assessed against ground truth values, represented by the central elements in affect segments.

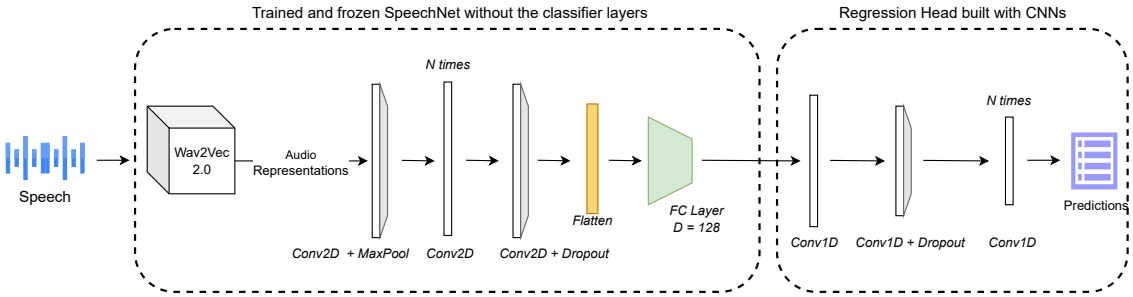


Figure 2.4: Trained SpeechNet, excluding classifiers, outputs from a fully connected layer of $D = 128$. A regression head on top predicts affect values, focusing on the middle of the affect contour.

2.3 Experimental Evaluation

The proposed architecture is implemented using the PyTorch library. Experiments are conducted exclusively on the CPU. The Adam optimizer is employed for training with a learning rate set at 0.001. Training occurs in batches of 256 samples for a total of 50 epochs based on the preliminary experimentation on RECOLA. In addition to that, we include early stopping criteria to halt training if consecutive epochs fail to exhibit improvement beyond a specified margin. Hyper-parameters are fine-tuned throughout the experimentation process. Detailed information regarding the network's layers and learnable parameters are given in their respective sections.

Chapter 3

RESULTS

In this chapter, we will initially give information about the dataset employed in this study, depicting the preprocessing procedures applied to the data before feeding it into the neural network. After presenting the data, this chapter will introduce the empirical outcomes acquired during the research and will comprise two sections. In the initial part, emphasis will be placed on tracing the outcomes associated with the collaborative neural network architecture. This involves an initial examination of the networks' classification performance to predict the actual affect classes obtained as affect network output from intermediate layers. Additionally, the presentation will extend to explain the resulting impact of the classification task on the data during training, highlighting the alterations relative to their respective states before the network is applied to data. The underlying rationale for adopting this approach lies in extracting information to comprehend the beneficial influence of the joint network architecture on the regression task. An analysis of how the generated clusters evolve and align during the classification task is investigated. Comparative studies of cluster characteristics will yield insights into the effectiveness of the joint network architecture. In the latter part of the chapter, attention will be directed towards revealing observations about the regression performance of the secondary network, namely the regression network superimposed onto the trained speech network to predict affect contour values, which are extracted from the middle of the affect segments. The results will be presented alongside relevant literature, employing benchmark metrics widely recognized for cross-evaluating the efficacy of the proposed network.

3.1 Dataset Overview and Preprocessing

In this study, we conduct experiments using two databases, RECOLA and USC CreativeIT, which are commonly employed in emotion recognition tasks utilizing speech, visual, and motion data. Since our focus is specifically on speech emotion recognition within the scope of this research, so we will only be utilizing the speech data from these databases for our experiments.

The RECOLA database [Ringeval et al., 2013] is widely recognized as a prominent resource extensively employed in various multimodal emotion recognition tasks, featuring prominently in AVEC challenges across different years, including [Valstar et al., 2016] and [Ringeval et al., 2018]. It encompasses a substantial 9.5 hours of multimodal recordings, encompassing audio, visual, and physiological data (specifically, electrocardiogram and electrodermal activity). These recordings capture online dyadic interactions involving 46 French-speaking participants collaboratively engaging in a task. The audio subset of the dataset, pivotal for our study, comprises 27 five-minute speech utterances divided into training, development, and test sets, each paired with corresponding affect contours. The audio modality and its corresponding affective contours underwent thorough annotation by six gender-balanced annotators, who annotated activation and valence values at a frame rate of 40 milliseconds. This study uses the mean value derived from these six annotations as the benchmark for our evaluation.

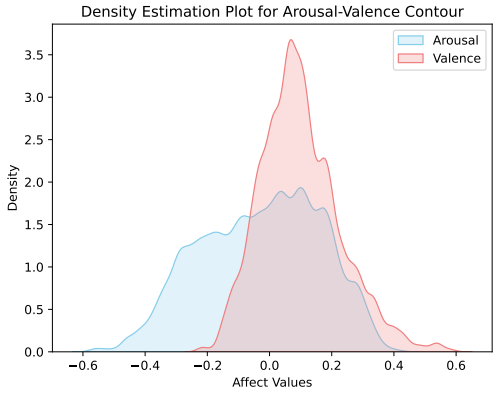
USC CreativeIT [Metallinou et al., 2010, Metallinou et al., 2015] database, a multimodal database of theatrical improvisations, features pairs of 3.5-minute interplays performed by 16 actors to simulate dyadic interactions, with annotations provided by 3 raters at a frequency of 60 Hz. Each interaction within the USC CreativeIT database contains two recordings, capturing audio and motion data, enabling the extraction of speech and body motion features. Once recordings lacking labels for any arousal and valence contours are removed, a training dataset is created from 40 recordings. Ground truth annotations are derived from mean annotator ratings for each recording.

Figure 3.1 below depicts the databases' affect contour values for arousal and

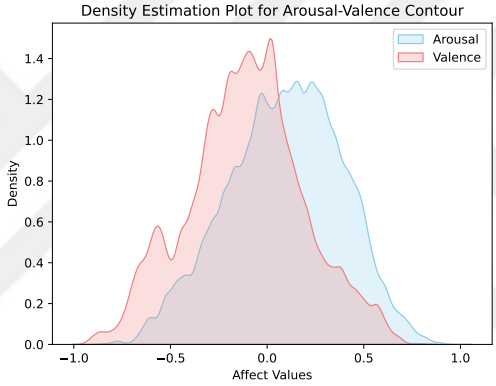
valence using kernel density estimation (KDE). Examining the graph reveals that arousal values are spread across the entire range for the RECOLA database, resembling a normal distribution. On the other hand, the valence values are concentrated predominantly around the median level in the same fashion, indicating a denser population in that range. The distribution of labels in RECOLA is narrower in both affective dimensions than CreativeIT, possibly due to its purely spontaneous nature. While actors contributing to the CreativeIT database were not restricted in their choice of words for interaction, they were bound by the general scenario and assigned roles. These conditions may have resulted in more characteristic emotions in CreativeIT.

3.1.1 Data Preprocessing

In the preceding section, we outlined our approach to averaging annotations provided by independent annotators. Before data input, both datasets undergo several procedural steps. The training, development, and test set each comprises nine five-minute speech data segments, concatenated to yield a total of 45 minutes of speech data for the RECOLA dataset. Similarly, a total of 40 recordings with a 3.5-minute duration are utilized for the USC CreativeIT dataset. The corresponding affect contour values are aggregated into a one-dimensional vector. The procedure entails extracting 2-second speech segments from the initial combined speech data, shifting the focal window by 1 second and pulling another 2-second speech segment until the entire speech dataset is processed. Affective vectors are generated similarly, with each 2-second speech signal corresponding to a 50-unit-long affect vector for the RECOLA dataset. To align with the speech data, the first 50-unit affect vector is extracted from the combined affect contour, and 25 units are then shifted in the window to obtain successive samples until the entire spectrum is scanned. Regarding the USC CreativeIT dataset, the annotation rate (60 Hz) differs from that of the RECOLA dataset, resulting in different affect vector lengths for the 2-second audio sequences. We opt for a 2-second audio sequence for the USC CreativeIT dataset to ensure consistency and facilitate cross-dataset comparison in our experiments. Con-



(a) RECOLA Database



(b) USC CreativeIT Database

Figure 3.1: Plots for affect contour distribution of databases using KDE

sequently, the corresponding affect contour is generated using the 60 Hz annotation rate, resulting in a length approximately 2.4 times longer than that of the RECOLA dataset.

During experimentation, we will exclusively display the model parameters for the RECOLA dataset, as there are no significant differences between the specific models for the datasets, except for the input shapes received by the model for the affect contour.

3.2 Experimental Results of the Classification Task

As outlined in the methodology chapter, the core of the joint neural network comprises two independent neural networks. The initial network is the speech network, and the second is the affect network, responsible for accurately categorizing affect clusters formed through its representative capacity in the intermediate layers. This section presents the classification outcomes associated with the speech network, given our primary objective of predicting affective states based on speech. Essential benchmark classification metrics, such as precision, recall, and F1 score, will be provided, delivering distinct perspectives for a comprehensive evaluation of the classification network. Beyond the classification performance of the speech network, we will enumerate outcomes related to the clusters and the distinctive features that set them apart. These results primarily examine the effectiveness of cluster generation during training. Eventually, the comparative analysis undertaken will clarify the reciprocal influence between the classification and clustering processes.

Regarding the classification outcomes, we will initially present the classification outcomes introduced at the beginning of this section. Subsequently, we will introduce a set of graphs illustrating the orientation of cluster classes based on their magnitudes and characteristic patterns. These graphs aim to provide a clearer understanding of the impact of the training procedure compared to the initial states when no network effect is applied. Therefore, this section presents the classification performance of the speech network within the joint neural network and examines how the joint neural network influenced the affect contour and its characteristics.

To determine the optimal cluster count for arousal and valence values, we charted the within-cluster sum of squared errors (WSS) against varying cluster numbers, as illustrated in Figure 3.2. Upon analysis, we discovered that the ideal cluster count for arousal and valence values is $K = 4$, indicated by a distinct elbow-like pattern around that value. Utilizing this finding, we proceeded with experiments.

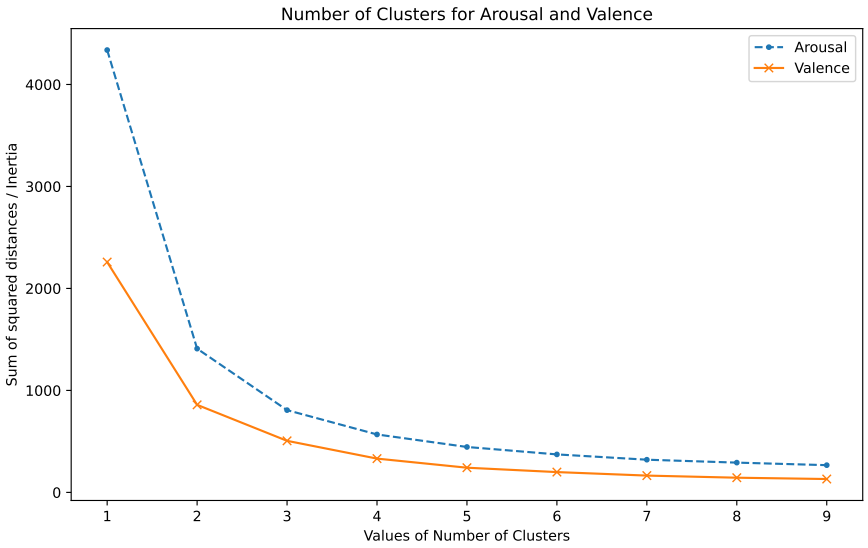


Figure 3.2: The plot of WSS concerning the number of clusters based on the RECOLA dataset

3.2.1 Classification Results

This section will initially present quantitative outcomes obtained during the experimentation for various metrics. To effectively demonstrate the impact of the joint neural network, we also conduct training on a model that, in contrast to the joint architecture, initially clusters the affect contour into distinct groups and does not utilize the affect network of the joint model. Instead, we exclusively train the speech network to establish a mapping from speech to affect clusters. The objective is to comprehend the results of a one-sided network at the baseline. In this phase, we maintain the structure of the speech network but exclude the affect network. For the baseline model, the affect segments with 50-unit-long vectors are used for one iteration of clustering using k means clustering. Another baseline model is also created by aggregating 50-unit-long vectors into scalar values. Then, these aggregated values are clustered instead of clustering the affect contour vectors. Consistently, because the CreativeIT dataset is annotated at a rate of 60Hz, the generated affect vector is about 2.4 times longer than its counterpart. The results of the baseline models,

where the affect clustering network is excluded, and the mean baseline model, where the clustering is applied to the mean values of affect contour, are included in the classification result table as a comparison point. Table 3.1 and Table 3.2 below display the results of the proposed method, baseline method, and mean cluster method for comparison while clustering the arousal and valence contours into four classes. As the datasets are already split into training and test sets, we use them accordingly, and the results are obtained from the test set.

Table 3.1: Classification results for joint and baseline models for RECOLA dataset

Model	**Arousal Values**				**Valence Values**			
	A	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**
Joint Model	0.83	0.90	0.79	0.84	0.71	0.78	0.72	0.75
Baseline Model	0.34	0.38	0.36	0.37	0.29	0.32	0.31	0.31
Mean Baseline Model	0.24	0.26	0.28	0.27	0.25	0.24	0.27	0.25

Table 3.2: Classification results for joint and baseline models for CreativeIT dataset

Model	**Arousal Values**				**Valence Values**			
	A	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**
Joint Model	0.66	0.68	0.63	0.65	0.52	0.54	0.53	0.53
Baseline Model	0.29	0.32	0.31	0.31	0.27	0.28	0.28	0.28
Mean Baseline Model	0.21	0.22	0.2	0.21	0.2	0.21	0.2	0.2

As depicted in Table 3.1 and Table 3.2, the proposed joint model demonstrates superior performance across all metrics compared to the baseline model. The table also illustrates that the baseline model exhibits marginal improvement over random selection. The mean baseline model is the worst-performing in all three network structures. As the affect value distribution is close to a normal distribution, tak-

ing the mean of the affect contour worsens the general prediction capacity of the model. When comparing the results across datasets, it becomes apparent that the RECOLA dataset consistently outperforms the CreativeIT dataset across all experimental models.

Examining the characteristics of the affect clusters after training is crucial, as it directly impacts classification performance. Our focus is on understanding the generation of clusters, mainly by analyzing the outputs of the intermediate layer of the affect network dedicated to clustering affect contours. The latent representations for clustering, with a dimensionality of $D = 8$, undergo a dimension reduction process from an initial 50-dimensional vector for RECOLA. Given that k-means clustering serves as a form of vector quantization, mapping observations to the nearest cluster centroid, our interest lies in determining whether the partitioned data displays distinctive characteristics beyond simple quantization into specific levels within the space. We aim to explore the cluster centroids visually.

The following pair of graphs will characterize the cluster centroids. The first graph represents centroids generated by the affect network before the start of training, while the second one depicts centroids after the completion of training in the joint architecture. The goal is to analyze these graphs in order to distinguish the impact of the joint network procedure on the formation of clusters.

The magnitudes of cluster vector elements are represented on the y-axis, while the x-axis corresponds to the indices in the affect vector.

As provided in Figure 3.4, distinct slopes and magnitudes characterize the cluster centroids, allowing for clear differentiation. To complement this visual representation, we will incorporate the initial cluster centroids generated for the identical network before the initiation of training, as illustrated in the subsequent graph.

Upon analyzing the graphs, it becomes apparent that the initial cluster centroids, established through k-means clustering, encode magnitude level information by associating samples with the nearest quantized levels. However, after the training phase, these centroids transform, displaying varied magnitude levels and distinct patterns compared to neighbouring centroids. This shift indicates a departure from the initially fixed magnitude levels. Furthermore, the y-axis magnitude range for

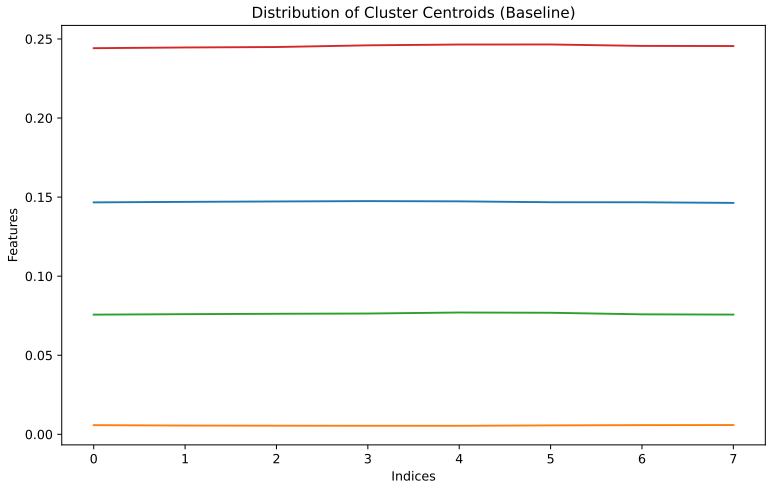


Figure 3.3: 8-dimensional arousal cluster centroids of RECOLA dataset before the start of the joint architecture training process

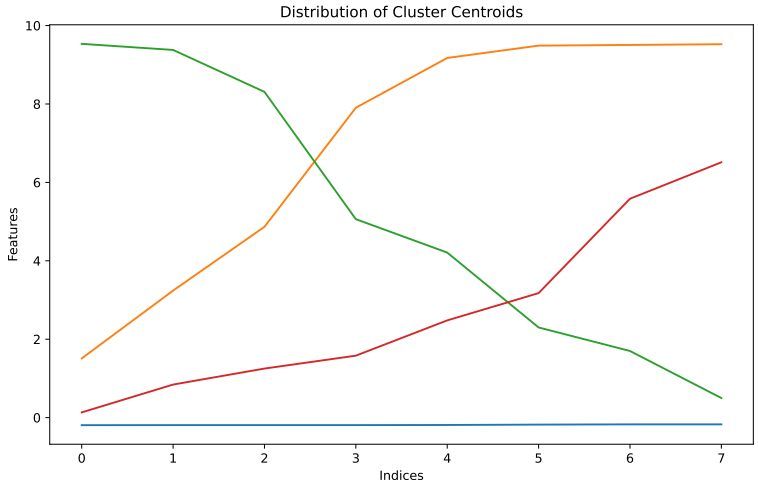


Figure 3.4: The plot for 8-dimensional cluster arousal centroids of RECOLA after training

the baseline centroids originally spans from 0 to 0.25 values, but post-training, this range expands to encompass values between 0 and 10.

The observed increase in centroid range during training implies an improvement in resolution. In contrast to the initial graph, where clusters are differentiated by

quantization linked to magnitudes, the trained clusters exhibit a change in nature, showcasing diverse patterns rather than encoding magnitude levels.

We identified clusters for the batches to assess the clustering mechanism's impact on 50-unit-long affect vectors. Grouping vectors by their cluster labels, we created 50-unit-long aggregated vectors for each label. Plotting them on a shared graph revealed insights into how the clustering mechanism influences affect vector representation in two-dimensional space.

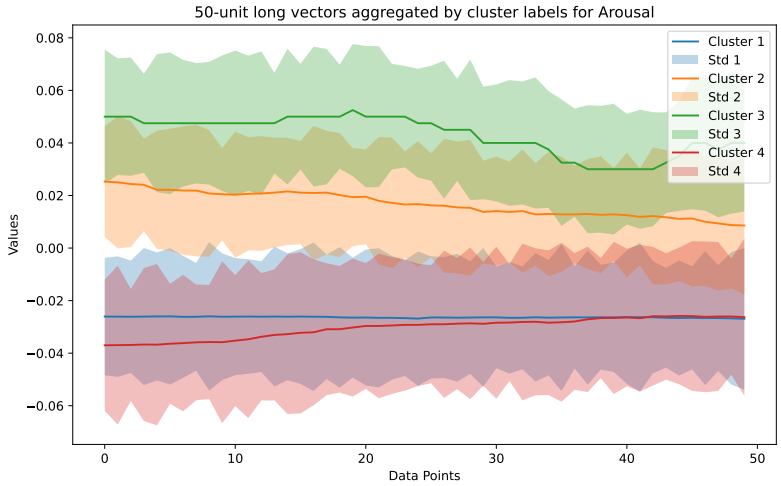


Figure 3.5: The plot for mean cluster vectors with standard deviations for 4-class clustering and classification network of RECOLA arousal contour

The results for the classification section are presented employing 4-class cluster targets, which are generated utilizing 8-dimensional latent representations from the affective clustering network. On the other hand, the mapping from speech to affect clusters is based on the 128-dimensional fully connected layer from the speech network. The choices for the presented results are obtained by a brute search fashion where we change these parameters number of clusters (K), the dimensions of the vectors that generate cluster vectors (D) and the latent representations taken from the speech network) to obtain the result that gives the higher scores for classification tasks. Besides, when assessing the contributing parameters to the classification network, we also consider the effects of these parameter choices on the continuous

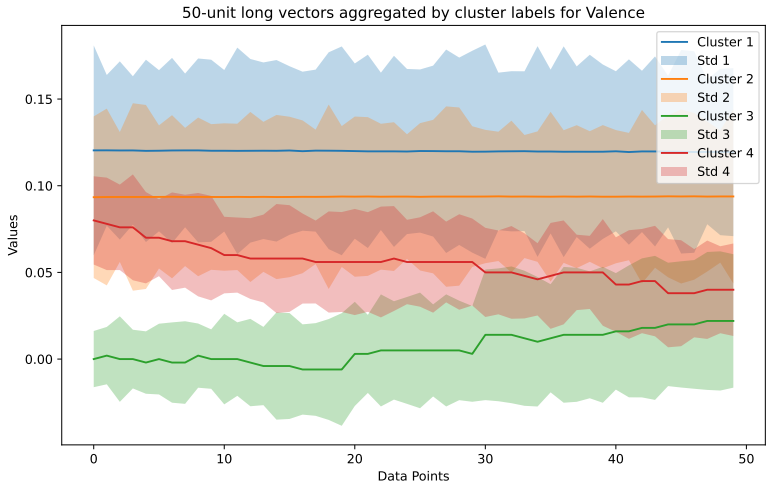


Figure 3.6: The plot for mean cluster vectors with standard deviations for 4-class clustering and classification network of RECOLA valence contour

recognition task. Generally, the setup with a 4-class cluster network outperforms the other setups with different cluster numbers. This finding is obtained for all the setups when only the cluster number is changed, and all other controlling parameters are left unchanged. The first step in this approach is to specify the clustering number. Then, we optimize parameters to improve classification and continuous regression performance. The number of clusters is found to be the most critical parameter for improving performance. Once the number of clusters is determined, our focus shifts to the dimension of the affect vectors undergoing clustering to predict affect clusters. During this stage, we observe that cluster vectors remain relatively unchanged towards the end of the training, and vectors with dimensions close to affect segments of 50 units provide little information. As a result, in the subsequent iterations, we progressively select vectors with lower dimensions from the deeper layers of the affect network. This reveals a correlation between vector dimensionality and cluster vector convergence from baseline. Ultimately, we find that 8-dimensional vectors perform best, and our experiments are based on these 8-dimensional vectors. In Figure 3.5 and Figure 3.6, it's evident that there's minimal overlap among the clusters, particularly those displaying ascending and descending trends, based on

their standard deviations. This indicates that the clusters generated can be distinguished from each other based on the distribution of their magnitude levels across the arousal and valence contours. In Figure 3.5 and Figure 3.6, it's noticeable that there's a slight overlap among the clusters, particularly those with ascending and descending trends, based on their standard deviations. This suggests that the clusters generated can be differentiated from each other based on the distribution of their magnitude levels across the arousal and valence contours.

As part of further exploration within the classification framework, we conducted an additional experiment to examine the relationship between the generated clusters for arousal and valence contours. To visualize this relationship, we created a heat map. In this heat map, we denoted their respective arousal and valence clusters obtained during the training for a specific speech and affect segment. Given that these clusters exhibit distinct tendencies regarding magnitude levels (whether they increase, decrease, or remain stable), we anticipate observing specific patterns in their pairings. The resulting plot, depicted in Figure 3.7, illustrates this analysis.

Figure 3.7 shows dense pairings between increasing and decreasing arousal clusters and increasing and decreasing valence clusters. These pairings create nearly equal distributions across four combinations based on the changes in both contours. These areas on the two-dimensional arousal-valence plot correspond to the four quadrants that span the entire two-dimensional space. Additionally, a significant pairing occurs between an increasing trend in arousal and no change in valence. This suggests that changes in arousal are more perceptible to humans, while valence changes may be less distinct based on human annotations, an expected situation discussed in previous studies.

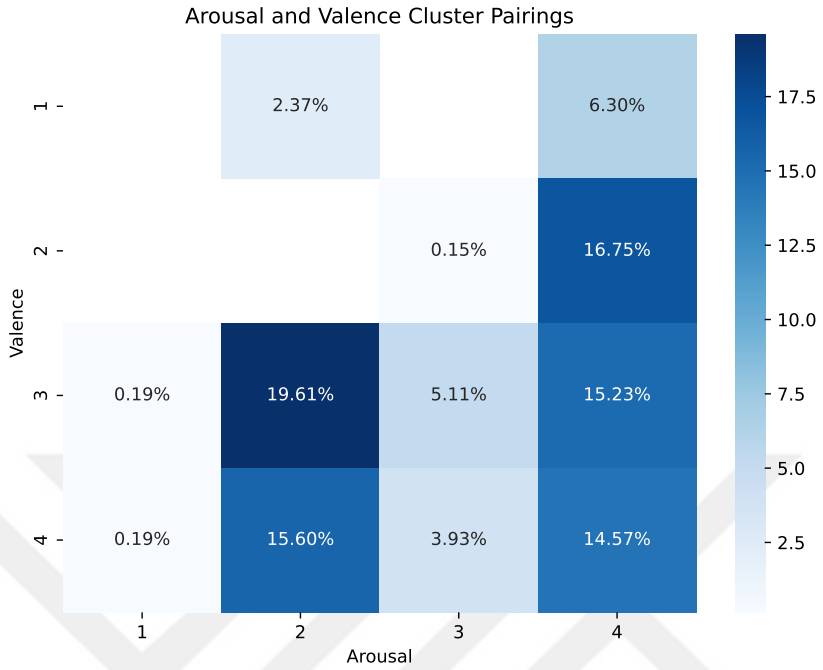


Figure 3.7: Percentages of cluster pairing for arousal and valence contours over the RECOLA dataset. Arousal clusters: 1 is stable, 2 and 3 decrease, and 4 increases. Valence clusters: 1 and 2 remain unchanged, 3 increases, and 4 decreases.

3.3 Continuous Emotion Recognition Results

This segment outlines the outcomes obtained from the regression task, as previously defined and expounded upon in the preceding sections. The assessment of the regression task primarily relies on the performance of the regression network executed on test data. Evaluation metrics for this section encompass benchmark regression criteria, including Root Mean Square Error (RMSE) and Concordance Correlation Coefficient (CCC). This deliberate selection facilitates a comprehensive comprehension of the data and allows the opportunity to compare the results with existing literature. The metrics chosen align with those recognized as benchmarks in research about continuous emotion recognition and related challenges, as observed in affective computing challenges such as AVEC [Ringeval et al., 2019] and AFF-WILD

[Kollias et al., 2018].

The subsequent part of this chapter will follow a structure where we first elaborate on the evaluation metrics employed in the regression task. We will then present our findings in tabular form, comparing them with the results reported in the existing literature.

3.3.1 Evaluation Metrics

The Root Mean Square Error (RMSE) serves as a metric for assessing the average magnitude of differences between observed values (y_i) and their corresponding predicted counterparts ($\hat{y}_i$). This measurement involves taking the square root of the mean of the squared differences between each observed and predicted value. In this context, n denotes the total number of data points, y_i signifies the observed value at position i, and $\hat{y}_i$ represents the predicted value at position i. In essence, RMSE values provide insights into the model's ability to fit the data effectively and make accurate predictions, considering the cumulative impact of numerous individual observations.

$$\text{Root Mean Square Error } \textit{(RMSE)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.1}$$

The Concordance Correlation Coefficient (CCC) combines Pearson's correlation coefficient (CC) with the squared difference between the means of two compared time series. In this context, ρ represents the Pearson correlation coefficient between two-time series: the prediction and reference. Moreover, σ_x^2 and σ_y^2 denote the variances of each time series, and μ_x and μ_y signify the mean values of each. Consequently, predictions exhibiting a strong correlation with the reference but experiencing a shift in values are subject to proportional penalization based on the magnitude of the deviation. In brief, concordance correlation coefficients evaluate both the strength and direction of linear relationships between pairs of variables, offering a holistic perspective that considers the interconnection between variables rather than examining individual relationships with one variable at a time.

$$\text{Concordance Correlation Coefficient } (CCC) = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{3.2}$$

3.3.2 Experimental Results

The proposed network undergoes experimentation using unseen test data generated by the same human speaker as the training set. As previously mentioned, during experimentation, we specifically chose the middle element from each affect segment created in the data preparation phase. While existing literature typically compares networks using common test data, our approach diverges from this precedent. Nevertheless, we contend that predicting the middle elements of the segments still yields meaningful information about the network's performance.

We utilize recent publications and the baseline performance established in the [Valstar et al., 2016] challenge to compare the CCC metric. The test sets in the challenge are created using a hop size of 400 ms between consecutive samples, and they are concatenated to form the final test set. The experimentation mimics this by making a hop size of 2000 ms between the middle points of the consecutive affect segments. This will provide a decent comparison point for the network. The initial publication, denoted as [Khorram et al., 2017b], explores two distinct architectural strategies for capturing temporal dependencies within a sequence of acoustic features. The first method employs dilated convolutions with filters featuring varying dilation factors, while the second method utilizes downsampling/upsampling networks with a sequence of convolutions to downsample the signal. The subsequent publication, identified as [Han et al., 2017], introduces a soft prediction methodology to augment automatic emotion recognition systems for more refined emotion analysis. This approach utilizes BLSTM-RNN [Graves and Schmidhuber, 2005] regressors to predict emotional states. In [AlBadawy and Kim, 2018], a discretized and continuous neural network is trained end-to-end to obtain a final prediction utilizing deep LSTM networks. Table 3.3 presents quantitative outcomes for continuous emotion recognition tasks using the RECOLA dataset, compared with findings from existing literature.

Table 3.3: CCC values comparison for arousal and valence values for RECOLA dataset experimentations

Model	**Arousal**		**Valence**	
	dev	**test**	**dev**	**test**
[Valstar et al., 2016]	0.796	0.648	0.455	0.375
[Han et al., 2017]	0.858	0.682	0.563	0.448
[Khorram et al., 2017b]	0.867	0.684	0.592	0.502
[AlBadawy and Kim, 2018]	0.868	0.693	0.623	0.530
Our method	0.821	0.662	0.580	0.494

In addition to the comparison outlined in the previous table, where we assess our work concerning existing literature, we also examine the results of the regression task against our baseline. Our baseline model is constructed similarly, involving clustering on the entire dataset without undergoing the joint neural network procedure or incorporating joint clustering and classification training. Consequently, the baseline model lacks the information gain our affect clustering network provided. The remaining structure of the network stays unchanged for the classification part, with the speech network trained to classify affect segments. The regression head is also placed on the same intermediate layer of the speech network for continuous emotion recognition tasks. This controlled experiment enables us to evaluate the contribution of the joint network structure to the regression task. Following the completion of the procedure, we observe that the CCC values for both arousal and valence contours are lower than their counterparts for both the development and test sets.

Correlation results for both the proposed network and those obtained from the existing literature are presented in the previous section. The Concordance Correlation Coefficient (CCC) is valuable when assessing the correlation between two continuous variables. It provides a robust measure of concordance, considering the standard deviations and means of both sets of continuous variables. A comprehen-

Table 3.4: RMSE values comparison for arousal and valence values

Model	**Arousal**		**Valence**	
	dev	**test**	**dev**	**test**
[Povolny et al., 2016]	0.114	0.141	.114	.114
[Le et al., 2017]	0.103	0.143	0.113	0.116
[Khorram et al., 2017a]	0.100	0.137	0.107	0.117
[AlBadawy and Kim, 2018]	0.868	0.693	0.623	0.530
Our method	0.107	0.142	0.111	0.114

sive understanding of individual sample-level deviations from the target value can now be obtained by shifting our focus to another metric, the Root Mean Square Error (RMSE). For this aim, in Table 3.4, we present the results of the proposed method along with the recent literature.

After presenting and comparing the experimental findings with existing literature on the RECOLA dataset concerning the continuous emotion recognition task we defined, we now turn to the results for the CreativeIT dataset we utilized alongside RECOLA. Using the concordance correlation coefficient (CCC) metric as our primary comparison metric against prior literature, we will compare the results of our proposed model with those presented in the paper by [Fatima and Erzin, 2021] as it includes CER results based on CCC values. As Section 1.2 mentions, the paper introduces various architectures incorporating dyadic context and temporal correlations while employing a multimodal approach that includes speech modality. We will compare our proposed method with three different architectures from the paper. Firstly, the CER-CNN model, single subject CNN structure, serves as a baseline, is utilized as an affect context estimator and is constructed using convolutional layers. The second and third architectures extend the baseline CER-CNN model. The second one, called the Dyadic FoA Model, incorporates the fusion of cross-subject affect, while the third, the Dyadic FoM Model, integrates the fusion of cross-subject feature maps. Both architectures establish dyadic affect context and

employ a two-stage CER network built on the CER-CNN model.

Table 3.5: CCC values comparison for arousal and valence values for CreativeIT dataset experimentations

Model	**Arousal**	**Valence**
CER-CNN (Baseline)	0.1598	0.0421
Dyadic FoA Model	0.1734	0.0560
Dyadic FoM Model	0.1447	0.0530
Our Method	0.1623	0.0537

Table 3.5 compares the concordance correlation coefficients. According to the results, one of the first conclusions is that arousal predictions outperform the valence values in our proposed model and previous studies. This trend may stem from the fact that arousal can be more precisely depicted through speech acoustic signals. Our approach demonstrates competitive outcomes in audio-only scenarios compared to the previous study. Although the baseline model already exhibits competitive performance and delivers decent continuous emotion recognition (CER) results, our method outperforms the baseline and the architecture utilizing cross-subject feature maps. Finally, our proposed method performs marginally worse than the model employing cross-subject affect.

Chapter 4

CONCLUSION

In the field of emotional analysis, extensive research has focused on predicting emotional attributes (regression) and emotional categories (classification) from speech data. However, this remains an active study area due to the inherent challenges in representing and quantifying emotions across different modalities, particularly speech. To address this, we propose a novel semi-supervised learning method. Our approach aims to identify clusters of temporal affect contours that can be effectively predicted from speech representations.

We evaluated our method using the RECOLA dataset and partially on the USC CreativeIT dataset. The results demonstrate that the learned affect contour clusters, focusing on speech predictability, achieve promising performance. Specifically, in our four-class classification tasks, the learned affect clusters achieved F1 scores of 0.84 for arousal and 0.75 for valence. Furthermore, we observed dominant trends in the movement of arousal-valence contours across the four quadrants of the arousal-valence (AV) plane.

Additionally, this thesis explores framing a regression challenge as a joint classification and clustering problem. Rather than treating continuous variables in isolation, we conceptualize them as part of higher-level clusters or classes that share common properties and exhibit similar behaviours. This involves merging consecutive individual annotations to create constant-length vectors. By employing networks that map affect contours to form affect classes and speech content to affect clusters, we establish training for a joint network—analogous to self-supervised training in neural networks.

While both datasets undergo similar treatment in the study, they exhibit differences for several reasons. Firstly, the data collection environments vary: the RECOLA database involves speakers engaging in natural conversations for a prede-

termined duration, whereas CreativeIT prompts participants to adhere to a standard conversation script. Additionally, annotators remain consistent throughout sessions for RECOLA, whereas they vary for CreativeIT. Another crucial difference lies in the annotation rates, which significantly differ between the datasets. Initially, the higher annotation rate in CreativeIT may suggest a more comprehensive data representation. However, results indicate that despite CreativeIT containing more data points, it does not necessarily yield better Continuous Emotion Recognition (CER) performance. Consequently, experiments commence with RECOLA, maintaining a consistent setup across both databases for comparison. An alternative setup may be more effective for datasets with higher annotation rates, capturing context information more accurately within shorter speech and affect contour segments.

The resulting vectors are then clustered into a finite number of classes, transforming the regression task into a classification domain. These classes' simultaneous clustering and classification yield refined cluster centroids that are distinguishable from neighbouring ones. To assess the advantages of treating the problem as a classification task, we use the trained speech network as a feature extractor for the defined regression task. We reintroduce the regression problem to predict single human-annotated affect values.

Through experimentation, we observe that the joint clustering and classification network enhances the regression task when the speech network is utilized as a feature extractor for the regression head, producing results that are competitive with those found in the existing literature. The novelty of our approach lies in its departure from quantization-based methods seen in some works, where the magnitudes of affect contour elements dictate the strategy. Instead, our approach avoids imposing strict quantization on the affect contour. Instead, it encourages a blend of discretization that considers not only the magnitude levels of the contour elements but also the relationships between neighbouring elements and the patterns they form, contributing to the creation of cluster classes.

The connection between consecutive affect elements serves as a normalization factor for individual annotations. This ensures that they are not interpreted in isolation but considered members of the classes to which they belong. Consequently,

class membership influences the dynamics of affect contour values, introducing a different perspective that differs from traditional quantization-based approaches found in the literature.

4.1 Future Directions

In this thesis, we examined the impact of quantization and the introduction of affect clusters on improving the performance of regression tasks while also achieving distinguishable classes that share common characteristics. As a future area of exploration, we aim to develop a comprehensive framework that allows us to establish a relationship between the number of clusters and their measurable properties, such as lengths, about the information gain transferred to the regression task. This framework would provide insights into how the formation of clusters influences the classification and regression performance of the problem. Additionally, it would help determine the minimum number of clusters and cluster vectors necessary to share the most common characteristics. This approach would facilitate a deeper understanding of how temporal information is encoded within the clusters, clarifying how much of this temporality can be utilized for clusters with specific lengths and varying total numbers of cluster classes. Another investigation route involves exploring the created clusters' real-world implications and decoding how to interpret these clusters in a manner perceptible to humans. Comparable to grouping dog and cat images into the same class for a vision problem without prior knowledge of their belonging to these specific classes, this exploration may necessitate quantitative and qualitative approaches to utilize the acquired information effectively.

In the context of this work, we are working with the affective arousal contour and the valence contour independently of each other. For future work, to see how the proposed network behaves when spanning the entire emotional space, we plan to work on 2-dimensional arousal-valence (AV) data by combining the affective contours. This would provide a holistic understanding and a direct relationship between the results of the network and affective emotional states.

BIBLIOGRAPHY

[Abdel-Hamid, 2020] Abdel-Hamid, L. (2020). Egyptian arabic speech emotion recognition using prosodic, spectral, and wavelet features. *Speech Communication*.

[Akash et al., 2016] Akash, S., Aschana, K., Abhijith, M., and Shuvalila, M. (2016). Speech based emotion recognition system. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 5(6):39–42.

[AlBadawy and Kim, 2018] AlBadawy, E. A. and Kim, Y. (2018). Joint discrete and continuous emotion prediction using ensemble and end-to-end approaches. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI '18, page 366–375, New York, NY, USA. Association for Computing Machinery.

[Andayani et al., 2022] Andayani, F., Theng, L. B., Tsun, M. T., and Chua, C. (2022). Recognition of emotion in speech-related audio files with lstm-transformer. In *2022 5th International Conference on Computing and Informatics (ICCI)*, pages 087–091.

[Arias et al., 2014] Arias, J. P., Busso, C., and Yoma, N. B. (2014). Shape-based modeling of the fundamental frequency contour for emotion detection in speech. 28(1):278–294.

[Baevski et al., 2021] Baevski, A., Hsu, W., Conneau, A., and Auli, M. (2021). Unsupervised speech recognition. *CoRR*, abs/2105.11084.

[Baevski et al., 2020] Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, abs/2006.11477.

[Bakhshi et al., 2022] Bakhshi, A., Harimi, A., and Chalup, S. (2022). Cytex: Transforming speech to textured images for speech emotion recognition. *Speech Communication*, 139:62–75.

[Burkhardt et al., 2005] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., and Weiss, B. (2005). A database of German emotional speech. In *Proc. Interspeech 2005*, pages 1517–1520.

[Busso et al., 2008] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, E. A., Provost, E. M., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

[Busso et al., 2007] Busso, C., Lee, S., and Narayanan, S. S. (2007). Using neutral speech models for emotional speech analysis. In *Proc. Interspeech 2007*, pages 2225–2228.

[Caponetti et al., 2011] Caponetti, L., Buscicchio, C. A., and Castellano, G. (2011). Biologically inspired emotion recognition from speech. *EURASIP journal on Advances in Signal Processing*, 2011:1–10.

[Chavhan et al., 2010] Chavhan, Y., Dhore, M., and Pallavi, Y. (2010). Speech emotion recognition using support vector machines. *International Journal of Computer Applications*, 1.

[Dahake et al., 2016] Dahake, P. P., Shaw, K., and Malathi, P. (2016). Speaker dependent speech emotion recognition using mfcc and support vector machine. In *2016 International Conference on Automatic Control and Dynamic Optimization Techniques (ICACDOT)*, pages 1080–1084.

[Devlin et al., 2018] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

[Ekman et al., 1987] Ekman, P., Friesen, W., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Krause, R., LeCompte, W., Pitcairn, T., Ricci Bitti, P., Scherer, K., Tomita, M., and Tzavaras, A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of personality and social psychology*, 53:712–7.

[Fatima and Erzin, 2021] Fatima, S. N. and Erzin, E. (2021). Use of affect context in dyadic interactions for continuous emotion recognition. *Speech Communication*, 132:70–82.

[Gao et al., 2017] Gao, Y., Li, B., Wang, N., and Zhu, T. (2017). Speech emotion recognition using local and global features. In *Brain Informatics: International Conference, BI 2017, Beijing, China, November 16-18, 2017, Proceedings*, pages 3–13. Springer.

[Graves and Schmidhuber, 2005] Graves, A. and Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610. IJCNN 2005.

[Han et al., 2017] Han, J., Zhang, Z., Schmitt, M., Pantic, M., and Schuller, B. (2017). From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty. *Proceedings of the 25th ACM international conference on Multimedia*.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

[Hinton et al., 2011] Hinton, G. E., Krizhevsky, A., and Wang, S. D. (2011). Transforming auto-encoders. In *ICANN*.

[Hou et al., 2020] Hou, M., Li, J., and Lu, G. (2020). A supervised non-negative matrix factorization model for speech emotion recognition. *Speech Communication*, 124:13–20.

[Huang and Bao, 2019] Huang, A. and Bao, P. (2019). Human vocal sentiment analysis.

[Iliou and Anagnostopoulos, 2009] Iliou, T. and Anagnostopoulos, C.-N. (2009). Comparison of different classifiers for emotion recognition. In *2009 13th Panhellenic Conference on Informatics*, pages 102–106.

[Issa et al., 2020] Issa, D., Fatih Demirci, M., and Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59:101894.

[Jannat et al., 2018] Jannat, S. R., Tynes, I., Lime, L., Adorno, J., and Canavan, S. (2018). Ubiquitous emotion recognition using audio and video data. pages 956–959.

[Jin and Han, 2010] Jin, X. and Han, J. (2010). *K-Means Clustering*, pages 563–564. Springer US, Boston, MA.

[Kerkeni et al., 2019] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M. A., and Cleder, C. (2019). Automatic speech emotion recognition using machine learning.

[Khorram et al., 2017a] Khorram, S., Aldeneh, Z., Dimitriadis, D., McInnis, M. G., and Provost, E. M. (2017a). Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition. *CoRR*, abs/1708.07050.

[Khorram et al., 2017b] Khorram, S., Aldeneh, Z., Dimitriadis, D., McInnis, M. G., and Provost, E. M. (2017b). Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition. *CoRR*, abs/1708.07050.

[Khorram et al., 2021] Khorram, S., McInnis, M. G., and Provost, E. M. (2021). Jointly aligning and predicting continuous emotion annotations. *IEEE Trans. Affect. Comput.*, 12(4):1069–1083.

[Kishore and Satish, 2013] Kishore, K. K. and Satish, P. K. (2013). Emotion recognition in speech using mfcc and wavelet features. In *2013 3rd IEEE International Advance Computing Conference (IACC)*, pages 842–847. IEEE.

[Kollias et al., 2018] Kollias, D., Tzirakis, P., Nicolaou, M. A., Papaioannou, A., Zhao, G., Schuller, B. W., Kotsia, I., and Zafeiriou, S. (2018). Deep affect prediction in-the-wild: Aff-wild database and challenge, deep architectures, and beyond. *CoRR*, abs/1804.10938.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

[Le et al., 2017] Le, D., Aldeneh, Z., and Provost, E. M. (2017). Discretized continuous speech emotion recognition with multi-task deep recurrent neural network. In *Interspeech*.

[Lin and Busso, 2024] Lin, W.-C. and Busso, C. (2024). Deep temporal clustering features for speech emotion recognition. *Speech Communication*, 157:103027.

[Lin et al., 2021] Lin, W.-C., Sridhar, K., and Busso, C. (2021). Deepemocluster: a semi-supervised framework for latent cluster representation of speech emotions. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7263–7267.

[Lotfian and Busso, 2019] Lotfian, R. and Busso, C. (2019). Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483.

[Mao et al., 2009] Mao, X., Chen, L., and Fu, L. (2009). Multi-level speech emotion recognition based on hmm and ann. pages 225 – 229.

[Martin et al., 2006] Martin, O., Kotsia, I., Macq, B., and Pitas, I. (2006). The enterface' 05 audio-visual emotion database. In *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pages 8–8.

[Metallinou et al., 2010] Metallinou, A., Lee, C.-C., Busso, C., Carnicke, S., and Narayanan, S. (2010). The usc creativeit database: A multimodal database of theatrical improvisation.

[Metallinou et al., 2015] Metallinou, A., Yang, Z., Lee, C.-C., Busso, C., Carnicke, S., and Narayanan, S. (2015). The usc creativeit database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations. *Language Resources and Evaluation*, 50.

[Milton et al., 2013] Milton, A., Roy, S. S., and Selvi, S. T. (2013). Svm scheme for speech emotion recognition using mfcc feature. *International Journal of Computer Applications*, 69(9).

[Neiberg et al., 2006] Neiberg, D., Elenius, K., and Laskowski (2006). Emotion recognition in spontaneous speech using gmms.

[Pichora-Fuller and Dupuis, 2020] Pichora-Fuller, M. K. and Dupuis, K. (2020). Toronto emotional speech set (TESS).

[Povolny et al., 2016] Povolny, F., Matejka, P., Hradis, M., Popková, A., Otrusina, L., Smrz, P., Wood, I., Robin, C., and Lamel, L. (2016). Multimodal emotion recognition for avec 2016 challenge. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, AVEC '16, page 75–82, New York, NY, USA. Association for Computing Machinery.

[Praseetha and Vadivel, 2018] Praseetha, V. and Vadivel, S. (2018). Deep learning models for speech emotion recognition. *Journal of Computer Science*, 14(11):1577–1587.

[Rieger et al., 2014] Rieger, S., Muraleedharan, R., and Ramachandran, R. (2014). Speech based emotion recognition using spectral feature extraction and an ensemble of knn classifiers. pages 589–593.

[Ringeval et al., 2018] Ringeval, F., Schuller, B., Valstar, M., Cowie, R., Kaya, H., Schmitt, M., Amiriparian, S., Cummins, N., Lalanne, D., Michaud, A., Ciftçi, E., Güleç, H., Salah, A. A., and Pantic, M. (2018). Avec 2018 workshop and challenge: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, AVEC'18, page 3–13, New York, NY, USA. Association for Computing Machinery.

[Ringeval et al., 2019] Ringeval, F., Schuller, B. W., Valstar, M. F., Cummins, N., Cowie, R., Tavabi, L., Schmitt, M., Alisamir, S., Amiriparian, S., Meßner, E., Song, S., Liu, S., Zhao, Z., Mallol-Ragolta, A., Ren, Z., Soleymani, M., and Pantic, M. (2019). AVEC 2019 workshop and challenge: State-of-mind, detecting depression with ai, and cross-cultural affect recognition. *CoRR*, abs/1907.11510.

[Ringeval et al., 2013] Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8.

[Shahin et al., 2022] Shahin, I., Hindawi, N., Nassif, A. B., Alhudhaif, A., and Polat, K. (2022). Novel dual-channel long short-term memory compressed capsule networks for emotion recognition. *Expert Systems with Applications*, 188:116080.

[Simonyan and Zisserman, 2015] Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. pages 1–14. Computational and Biological Learning Society.

[Szegedy et al., 2015] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

[Thayer, 1989] Thayer, R. E. (1989). *The Biopsychology of Mood and Arousal*. Oxford University Press USA.

[Valstar et al., 2016] Valstar, M. F., Gratch, J., Schuller, B. W., Ringeval, F., Lalanne, D., Torres, M., Scherer, S., Stratou, G., Cowie, R., and Pantic, M. (2016). AVEC 2016 - depression, mood, and emotion recognition workshop and challenge. *CoRR*, abs/1605.01600.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

[Wang et al., 2022] Wang, C., Ren, Y., Zhang, N., Cui, F., and Luo, S. (2022). Speech emotion recognition based on multi-feature and multi-lingual fusion. *Multimedia Tools and Applications*, 81.

[Wang et al., 2020] Wang, K., Su, G., Liu, L., and Wang, S. (2020). Wavelet packet analysis for speaker-independent emotion recognition. *Neurocomputing*, 398:257–264.

[Zamil et al., 2019] Zamil, A. A. A., Hasan, S., Jannatul Baki, S. M., Adam, J. M., and Zaman, I. (2019). Emotion detection from speech signals using voting mechanism on classified frames. In *2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST)*, pages 281–285.

[Zhang et al., 2021] Zhang, S., Tao, X., Chuang, Y., and Zhao, X. (2021). Learning deep multimodal affective features for spontaneous speech emotion recognition. *Speech Communication*, 127:73–81.

[Zhang et al., 2018] Zhang, Z., Han, J., Coutinho, E., and Schuller, B. W. (2018). Dynamic difficulty awareness training for continuous emotion prediction. *CoRR*, abs/1810.05507.

[Zhao et al., 2013] Zhao, X., Zhang, S., and Lei, B. (2013). Robust emotion recognition in noisy speech via sparse representation. *Neural Computing and Applications*, 24.