Jointly Aligning and Predicting Continuous Emotion Annotations

Soheil Khorram, Melvin G McInnis, Emily Mower Provost

Abstract—Time-continuous dimensional descriptions of emotions (e.g., arousal, valence) allow researchers to characterize short-time changes and to capture long-term trends in emotion expression. However, continuous emotion labels are generally not synchronized with the input speech signal due to delays caused by reaction-time, which is inherent in human evaluations. To deal with this challenge, we introduce a new convolutional neural network (*multi-delay sinc network*) that is able to simultaneously align and predict labels in an end-to-end manner. The proposed network is a stack of convolutional layers followed by an aligner network that aligns the speech signal and emotion labels. This network is implemented using a new convolutional layer that we introduce, the *delayed sinc layer*. It is a time-shifted low-pass (sinc) filter that uses a gradient-based algorithm to learn a single delay. Multiple delayed sinc layers can be used to compensate for a non-stationary delay that is a function of the acoustic space. We test the efficacy of this system on two common emotion datasets, RECOLA and SEWA, and show that this approach obtains state-of-the-art speech-only results by learning time-varying delays while predicting dimensional descriptors of emotions.

Index Terms—continuous emotion recognition, convolutional neural networks, delayed sinc layer, multi-delay sinc network.

◆

1 INTRODUCTION

EMOTIONS are complex and dynamic manifestations of internal human experience. Descriptions of emotion have often relied upon categories (e.g. happiness, anger, or disgust [1], [2]). However, these categories have substantial limitations due to the influence of cultural and other types of context [3]. As a result, the field of affective computing has increasingly adopted dimensional measures to quantify emotional expressions [4]–[14], most often by considering expression of valence (positive vs negative) and arousal (calm vs excited) [15]–[17]. These dimensional labels can be obtained by asking human raters to annotate data either statically, over units of input (e.g., segments or utterances) [18], [19], or continuously in time with a fixed sampling rate [20], [21].

Continuous assessments have the advantage of providing fine-grained information about the emotional state of the speaker as a function of time. One of the major challenges in recognizing continuous emotion labels (continuous emotion recognition) stems from annotators' reaction delay, which is defined as the amount of time it takes for annotators to sense acoustic events, understand them, and report the emotional labels [22]–[24]. Reaction delay is a convolutive noise that can shift continuous emotion annotations forward in time, causing a time difference between speech signals and the continuous emotion labels. This time difference depends on affective behaviors [23], which makes continuous emotion

recognition challenging. Therefore, in order to design an effective continuous emotion recognizer, it is crucial to compensate for the reaction delays.

The measure of human reaction delay and its relevance began in the modern era with the emerging focus on detailed experimental calculations in astronomy [25]. It was quickly established that reaction time is *individual dependent* [26], *stimulus dependent* [27], and *task dependent* [23], [28].

The importance of reaction delay compensation in continuous emotion recognition is widely recognized [22]–[24], [28]–[31]. There are two main approaches for handling this delay: (1) explicit compensation, in which researchers remove the delay in advance and then train emotion recognition systems [22], [23], [28]–[30]; these approaches assume that the delay compensation and the emotion recognition are independent; (2) implicit compensation, in which researchers build classifiers with large numbers of parameters to compensate for delay [32]–[34]. In this paper, we introduce a continuous emotion recognition system that is able to accomplish both goals: it compensates for annotators' delays while modeling the relationship between speech features and emotion annotations using classifiers with smaller numbers of parameters.

The proposed system is a convolutional network that is able to directly learn the annotators' delays. The network contains two components: (1) an emotion predictor and (2) an aligner. We train both components simultaneously in an end-to-end manner. The emotion predictor is a common convolutional network that models the relationship between acoustic features and emotion labels. The aligner network compensates for the annotators' delays using a new layer, the *delayed sinc layer*, which applies a learnable time-shift to a signal. The delayed sinc layer can be added to any network to compensate for misalignments between two signals.

The delayed sinc layer takes a one-dimensional signal and passes it through a shifted low-pass (sinc) filter. The

- *Soheil Khorram is a Research Fellow in the Department of Computer Science and Engineering, and the Department of Psychiatry, University of Michigan. E-mail: khorram.soheil@gmail.com*
- *Melvin G McInnis is the Thomas B and Nancy Upjohn Woodworth Professor of Bipolar Disorder and Depression, Department of Psychiatry, University of Michigan School of Medicine.*
  E-mail: mmcinnis@umich.edu
- *Emily Mower Provost is an Associate Professor in the Department of Computer Science and Engineering, University of Michigan.*
  E-mail: emilykmp@umich.edu

layer modifies its input by introducing a fixed delay, the amount of which is trainable through the back-propagation algorithm. We can handle variable delays by incorporating multiple parallel delayed sinc layers, each operating on a different region of the acoustic space. The shifted low-pass filter is also able to remove high frequency components of the input and generate a smooth output that is consistent with the slow moving nature of the emotion labels obtained from annotators (the importance of which is discussed in [22]).

The novelty of this work is the introduction and evaluation of the delayed sinc layer. We evaluate the proposed system on two publicly available continuously annotated emotion corpora: RECOLA [35] and SEWA [36]. We find that the proposed architecture obtains audio-only state-of-the-art performance on RECOLA. It also obtains audio-only state-of-the-art performance on SEWA when fused with another existing approach [37]. We demonstrate that a system that explicitly compensates for variable delay can use fewer parameters than one that implicitly compensates for delay through a large receptive field. We further demonstrate that a system that allows for flexibility in delay compensation outperforms systems that do not have this flexibility, noting that a single delayed sinc layer is not enough for modeling annotators' delay because this delay is non-constant. We find that arousal and valence prediction need at least 8 and 16 components, respectively, and delays over 7.5-seconds do not contribute to system performance. This suggests the importance of considering variability in delay (otherwise, the ideal number of components would be one). Finally, we investigate how reaction lag changes based on laughter, an emotionally salient event. We find that laughter regions of speech require smaller delays to be aligned with their emotion labels, compared to non-laughter regions.

2 BACKGROUND

As the main focus of this paper is to model and compensate for annotators' reaction delays, we provide a detailed overview for reaction delay compensation methods in Section 2.1. We also explain the state-of-the-art continuous emotion recognition systems developed for different datasets in Section 2.2. We compare our proposed network with two of these state-of-the-art systems in our experiments.

2.1 Reaction Delay Compensation

Reaction delay compensation techniques can be categorized into 2 groups: explicit and implicit.

2.1.1 Explicit Compensation

In this approach, the delay compensation and the emotion prediction are performed separately, which removes the need for the emotion predictor to handle delay. Researchers estimate the reaction delays by optimizing an alignment measure through a search algorithm. Different alignment measures have been proposed in different systems including mutual information and emotion recognition performance:

Mutual information – Mariooryad et al. estimated evaluator delay by maximizing the mutual information between the acoustic events and the continuous emotional labels.

Their experiments show that the mutual-information-based delay compensation technique can lead to seven percent relative accuracy improvement over baseline classifiers on the SEMAINE dataset [23], [28].

Recognition performance – The most common measure is the accuracy of the emotion recognition system. Trigeorgis et al. estimated the reaction time as a fixed value (between 0.04 and 10 seconds) that could be found through maximizing the concordance correlation coefficient between real and predicted emotion labels [30]. Huang et al. studied the effect of annotator's delay and introduced a number of multimodal emotion recognition systems based on an output associative fusion technique. They applied a temporal shift to each training sample to compensate for the annotation delay [22]. This temporal shift is performed by dropping first N emotion labels and last N input features. The value of the temporal shift, N, is tuned based on the development error during the training procedure. Their experimental results on AVEC 2015 confirm the importance of the delay compensation for continuous emotion recognition systems. They found that the best delays for arousal and valence are four and two seconds, respectively.

However, these approaches assume that the reaction delay is fixed for different acoustic events and that delay compensation and emotion prediction are independent and can be trained separately.

2.1.2 Implicit Compensation

In this approach, researchers leverage models that are able to compensate for delays while modeling the relationship between speech features and emotion labels. Different models have been used to this, including LSTM and convolutional networks:

LSTM network – Ringeval et al. [33] applied a long short-term memory (LSTM) network [38] with an analysis window to deal with simultaneous modeling of reaction delays and emotion labels. They found that predicting valence requires a longer analysis window compared to predicting arousal. Le et al. applied a multi-task bidirectional (B)LSTM network to model continuous emotion labels in a categorical time-dependent framework [34]. Their network is trained with a cost-sensitive cross-entropy loss function.

Convolutional network – In our previous paper [32], we employed two convolutional networks based on dilated convolutions [39] and downsampling/upsampling layers [40], [41] to predict continuous emotion labels. Both networks have large receptive fields that allow the networks to automatically shift the inputs forward in time and compensate for the reaction delay.

However, these implicit modeling approaches are not specifically designed to compensate for reaction delays. In this paper, we introduce the delayed sinc kernel, which is specifically designed to deal with delays in neural networks. Our experiments show that the proposed structure outperforms previous continuous emotion recognition systems.

2.2 Continuous Emotion Recognition

Continuous prediction of dimensional attributes has gained increasing attention over the last few years [42]–[45]. Several

competitions have been held in this research area, and different continuous emotion recognition systems have been proposed for each competition. For example, the audio/visual emotion challenge (AVEC) is a series of competition events aimed at comparing multimodal methods for recognizing emotion and depression patterns from audio, video, and physiological signals. In this section, we describe state-of-the-art emotion recognition systems developed for the AVEC competitions.

AVEC 2015 – In the winning submission of the AVEC 2015, He et al. [46] introduced an emotion prediction system with two phases. In the first phase, they obtained a set of initial predictions from each input modality using a BLSTM network. In the second phase, the initial predictions were smoothed with a Gaussian smoothing filter, and input into a multimodal BLSTM network for the final prediction of the affective states. The authors extracted a comprehensive set of $4,684$-dimensional low-level feature vectors from speech with the frame rate of 25 frames/s using both openSMILE [47] and YAAFE [48] toolkits, including loudness, zero crossing rate, spectral flux, Mel-frequency cepstral coefficients (MFCCs), and voicing related features (such as jitter, shimmer, logarithmic Harmonics-to-Noise Ratio (logHNR), etc). In our previous paper [32], we showed that a much smaller 40-dimensional MFB features with convolutional networks could be used to obtain state-of-the-art performance.

AVEC 2016 – In the best system submitted to the AVEC 2016 challenge, Brady et al. [49] employed a set of low-level audio features including MFCCs, shifted delta cepstral, and prosody features. The authors then trained a sparse coding technique over the low-level features to extract a set of higher-level audio features. The resulting higher-level features were then used to train linear SVRs for final continuous emotion recognition. In another successful system, Povolny et al. [50] extracted two sets of low-level audio features: (1) extended Geneva minimalistic acoustic parameter set (eGeMAPS) [51] and (2) bottleneck features obtained from intermediate representations of a DNN trained for automatic speech recognition (ASR) application. This DNN-based ASR is trained over an initial set of 24 log Mel filterbank (MFB) features and four different estimates of the fundamental frequency (F0). Povolny et al. used two methods for combining low-level features into high-level features that capture local contextual information: (1) simple frame stacking and (2) temporal content summarization by calculating statistics over local windows. The authors trained linear regressors to generate emotional labels from high-level statistics.

In our previous work [32], we showed that capturing long-term dependencies is beneficial for continuous emotion recognition. We studied two CNN-based architectures that are able to capture long-term temporal dependencies in a given sequence of acoustic features: dilated CNN and down-sampling/upsampling network. Dilated CNN uses a stack of convolutional kernels with varying dilation factors to capture long-term temporal dependencies. We showed that the dilated CNN outperforms previous systems, but it has an important problem: the output signals generated from this network undergo irregular changes between successive time steps. This noisy output is not consistent with the slow-moving emotion labels that are defined by annotators. We showed that a downsampling/upsampling network could be used to generate a smooth signal while considering the long-term dependencies for predicting emotion. It applies a series of convolutions and max-poolings to compress the input signal into a downsampled (low-resolution) signal. It then applies a series of transposed convolution[1] [52], [53] layers to upsample the compressed signal and generate the target emotional labels. This network achieved the best audio-only performance on the AVEC 2016 challenge. In this paper, we demonstrate that the proposed multi-delay sinc network outperforms the downsampling/upsampling network on the AVEC 2016 data.

AVEC 2017 – In the winning system of the AVEC 2017 [37], Chen et al. implemented a multi-task system that models and predicts multiple emotion labels simultaneously. They used IS10 [54] and Soundnet features (intermediate representations of the pretrained Soundnet network [55]) to train an LSTM-based continuous emotion predictor. LSTM network is able to alleviate the annotation delay problem and reduce the feature engineering efforts. In this paper, we show that fusing the predictions of the proposed network and the IS10-based system [37] yields the best audio-only performance on the AVEC 2017.

3 DATASETS AND FEATURES

In this section, we introduce the datasets, the features, the metrics, and the evaluation schemes used in this paper. We use two evaluation schemes, one based on the AVEC challenge (Section 3.2.1) and one using leave-one-speaker-out cross-validation scheme (Section 3.2.2).

3.1 Datasets

We use two publicly available datasets to conduct the experiments of this paper: (1) the remote collaborative and affective interactions (*RECOLA*) dataset [35] and (2) a subset of sentiment analysis in the wild (*SEWA*)[2] dataset. Both provide audio-visual recordings that capture spontaneous and naturalistic behaviors of subjects and are annotated with continuous emotion labels (arousal and valence values).

3.1.1 RECOLA

RECOLA was used in the multimodal affect recognition sub-challenge of AVEC 2016. It contains 27 samples of spontaneous and naturalistic interactions that were collected from 27 different French-speaking subjects. All samples are five minutes in length and include audio, video, electro-cardiogram (ECG) and electro-dermal activity (EDA). In this work, we focus only on the audio modality. The samples were partitioned uniformly into train, development and test sets by the organizers, nine per set.

The data were evaluated by six annotators (three female and three male), sampled at 25Hz. The evaluations include continuous assessments of valence and arousal. The six evaluation traces are fused into a single ground-truth using

---

1. In the literature transposed convolution is also known as deconvolution, upconvolution, fractionally strided convolution and backward strided convolution.

2. http://sewaproject.eu

the protocols described in the AVEC 2016 challenge [56]. Emotion labels are available only for the training and development sets of the data. Performance on the testing data is assessed by the organizers of the AVEC 2016 challenge.

3.1.2 SEWA

SEWA was used in the affect recognition sub-challenge of AVEC 2017. It is an audio-visual dataset of human-human interactions collected using common web-cams and microphones over the OpenTok API[3], an online platform for setting up a video call. Each recording includes two subjects discussing arbitrary aspects of a commercial that they have just viewed. The conversations range in length from 47-seconds to 3-minutes.

SEWA is a multi-lingual dataset, but the affect recognition challenge of AVEC 2017 used only the German-language recordings [36]. We also use the German-language subset in this paper. The subset contains 32 dyadic conversations (64 subjects in total), divided into three partitions (34 train, 14 development and 16 test). The data is split such that both subjects from a recording are in the same partition.

The data were evaluated by six annotators (three female and three male), sampled at 10 Hz. Again, the evaluations include continuous assessments of valence and arousal. We use the single gold standard label provided by the challenge [36]. As in RECOLA, the test labels are not released and performance of the system is assessed by organizers of the AVEC 2017 challenge.

3.1.3 Differences between RECOLA and SEWA

There are three important differences between the RECOLA and the SEWA datasets that should be considered in the design of our models:

1) In SEWA, each recording contains a conversation between two partners, one "target" and the other "non-target". The target partner is the partner whose emotions we aim to predict. In RECOLA, although the data were obtained from dyadic conversations, each recording includes only the audio from the target speaker.

2) In RECOLA, each recording has the same duration (5-minutes). In SEWA, the duration varies from 47-seconds to 3-minutes.

3) In RECOLA, the sampling rate of the emotion annotation traces is 25Hz. In SEWA, it is 10Hz.

3.2 Evaluation

The AVEC 2016 and 2017 challenges use the root mean square error (RMSE) and concordance correlation coefficient (CCC) metrics. RMSE is the standard error metric. CCC measures the agreement between two signals. It ranges from -1 to 1 and it is zero when two signals are uncorrelated from each other. It is defined by:

$$\text{CCC} = \frac{2\sigma_{y\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2}$$

where y and $\hat{y}$ are the sequences of ground-truth and predicted labels; μ_y and $\mu_{\hat{y}}$ are the mean of y and $\hat{y}$; σ_y^2

---

3. https://tokbox.com

and $\sigma_{\hat{y}}^2$ are the variance of y and $\hat{y}$; $\sigma_{y\hat{y}}^2$ is the covariance between y and $\hat{y}$.

3.2.1 AVEC evaluation scheme

This scheme follows the AVEC 2016 and 2017 guidelines. We train systems on the training/development partitions and assess performance on the test portion. We concatenate the output from each test recording into a single vector, which is then used to calculate the RMSE and CCC evaluation metrics. It is important to note that statistical tests cannot be performed in this setting because a single value is computed over all speakers and there are limited numbers of submissions allowed, thus precluding repeated assessments.

We train the network on the training partition, optimizing using the CCC metric over different sets of tuning parameters. We use the development partition to identify the set of tuning parameters that result in the highest performance. Finally, we use the identified network to generate labels for the held-out test data and submit the predictions to the organizers of the challenge. The organizers compute the final test evaluation metrics.

3.2.2 Leave-one-speaker-out evaluation scheme

The leave-one-speaker-out scheme addresses the limitation introduced by the AVEC challenge guidelines: the lack of ability to assess statistical significance. In this scheme, we perform leave-one-speaker-out cross-validation over the development speakers. We first train multiple networks with different hyper-parameters by maximizing average CCC values of the training samples. We then select hyper-parameters in a leave-one-speaker-out manner over the development set. We split the development set into speaker-specific folds (9 folds for AVEC 2016 and 16 folds for AVEC 2017). We leave out one fold for testing and choose the hyper-parameters using the remaining folds. We calculate metrics for each left out fold separately and then report the average as the final performance.

3.3 Features

Speech processing systems have relied upon a diverse set of spectral features, including linear prediction coefficients (LPC) [57], perceptual linear prediction (PLP) coefficients [57], mel-frequency cepstral coefficients (MFCC) [58], [59], mel-generalized cepstral coefficients (MGC) [60]–[63], and log mel-frequency bank (MFB) features [64], [65]. Emerging work has shown that emotion recognition systems can effectively use feature vectors composed solely of MFB features [7], [34] and that this small feature set can outperform much larger feature sets [32]. In this work, we extract 40-dimensional MFB features using the Kaldi toolkit [66] with a 25ms Hann window and 10ms frame shift. We apply speaker-specific z-normalization to reduce the speaker variability in the extracted features.

Feature vectors for the RECOLA dataset are created by concatenating every four successive MFB vectors to form a 160-dimensional vector; this creates a feature vector sampling rate that is consistent with the sampling rate of emotional labels [32], [67]. The SEWA dataset is more complicated to process because the dataset contains speech from both target and non-target speakers. We follow a

method, introduced by Chen et al. [37], that creates an 80-dimensional vector (40-dimensions per speaker). If a frame contains speech from the target speaker, the first half of the vector takes MFB values and the second half is zero and vice versa. We then concatenate every 10 consecutive feature vectors to again make the feature vector sampling rate consistent with the sampling frequency of emotion labels (10Hz). This process results in a sequence of 800-dimensional acoustic features.

4 PRELIMINARY EXPERIMENT

In this section, we set up an experiment that searches for an effective delay for RECOLA and SEWA in a manner similar to [23]. In later sections we will demonstrate that our networks are capable of learning these delays (see Section 7.3.1). The goal is to explicitly compensate for delay. We first time-shift all speech signals using a static shift; we then train a network that recognizes emotion labels from the delayed speech signals; finally, we calculate the leave-one-speaker-out CCC of the trained network; we repeat this procedure for different values of delay ranging from 0 to 6 seconds (step size of 400 milliseconds), and study the effect of the delay on the CCC results.

We employ a neural network consisting of a convolutional layer with one filter of length 2 seconds (50 frames on RECOLA and 20 frames on SEWA), followed by a $tanh$ activation unit, followed by a linear regression layer. The network is trained using the Adam optimizer over the CCC metric [68]. We select the number of training epochs and calculate the CCC values using the leave-one-speaker-out evaluation scheme explained in Section 3.2.

Figure 1 shows the CCC results with respect to the delay values. The results show that increasing the delay up to 2.4 seconds for arousal and 2 seconds for valence improves the leave-one-speaker-out CCC. It shows that synchronizing input and output (compensating for annotators' delay) is important in continuous emotion recognition. This result agrees with the previous findings (reported in [23]). We also find that when predicting valence, performance does not change given delays in the range of 2 to 3.2 seconds. This shows that we cannot find a unique value as the best estimate of the annotators' delay in detecting valence (i.e., the delay process follows a multi-modal distribution), which motivates the need to apply multiple delays for predicting continuous emotion labels.

This analysis presupposes that the reaction lag is a parameter of the network that can be tuned using validation data. This method of synchronizing input and output has several problems: (1) a separate network must be trained for any candidate value of the delay which is resource intensive; (2) the estimated lag must be a multiple of the sampling rate; (3) using this method, we cannot apply different delays to different regions of the acoustic features. In the next section, we introduce a method that can solve these problems.

5 METHODS

In this section, we introduce the *delayed sinc layer*: a convolutional layer that is able to apply a learnable shift to a given input signal. The delayed sinc layer is a time-shifted low-pass filter: it passes the frequency components that are lower
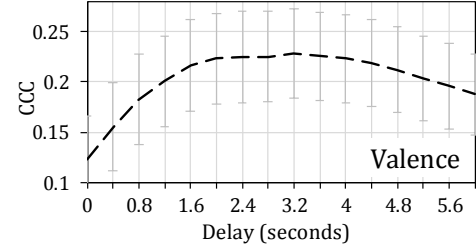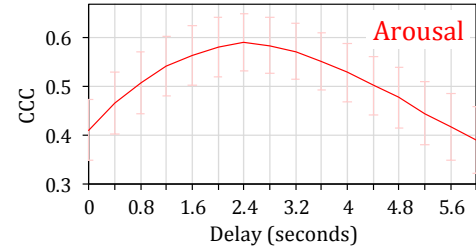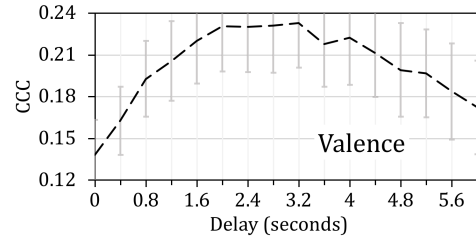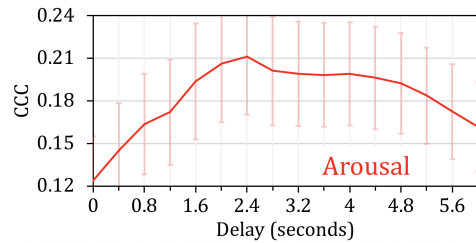
Fig. 1. Applying delay to acoustic features improves mean CCC for both arousal (solid red curve) and valence (dashed black curve). Error bars show standard deviation across different subjects in leave-one-speaker-out evaluation.

than a certain cutoff frequency and introduces a unique time-shift to its input. It allows us to align two signals in a neural network architecture.

We also demonstrate how the delayed sinc layer can be used in a continuous emotion recognition system. Our final network uses multiple delayed sinc layers and fuses the final outputs to compensate for signals that have multiple or time-varying delays.

5.1 Delayed Sinc Layer

Let $x(t)$ and $y(t)$ be continuous signals that are defined for acoustic features and emotion labels, at time t. The goal of continuous emotion recognition is to find a mapping $\mathcal{F}_{cer}$ that takes a sequence of acoustic features up to time t, $X(t) = [x(t') \ \forall \ t' \leq t]$, and estimates its corresponding emotion label at time t, $y(t)$:

$$y(t) = \mathcal{F}_{cer}(X(t)). \tag{1}$$

According to the preliminary experiment reported in Section 4, $x(t)$ and $y(t)$ are not synchronized. Therefore, $\mathcal{F}_{cer}$

should be written based on two simpler mappings: $\mathcal{F}_{syn}$ and $\mathcal{F}_{pred}$. $\mathcal{F}_{syn}$ performs the synchronization and $\mathcal{F}_{pred}$ models the relationship between acoustic features and the synchronized labels. In other words:

$$y(t) = \mathcal{F}_{cer}(X(t)) = \mathcal{F}_{syn}(\mathcal{F}_{pred}(X(t))). \qquad (2)$$

For the sake of simplicity, this section assumes that the synchronization can be done through applying a fixed delay τ (we will relax this assumption in Section 6). In this case, $\mathcal{F}_{syn}$ can be easily implemented through convolving $\mathcal{F}_{pred}(X(t))$ with a time-shifted dirac-delta function. Therefore, $y(t)$ can be written as:

$$y(t) = \mathcal{F}_{syn}(\mathcal{F}_{pred}(X(t))) = \mathcal{F}_{pred}(X(t)) * \delta(t - \tau), \quad (3)$$

where $*$ and δ represent the convolution operator and the dirac-delta function, respectively.

An effective approach to estimate τ is to learn it along with the parameters of $\mathcal{F}_{pred}$ through a gradient-based optimization technique. However, $\delta(t)$ is not a differentiable function whenever $t = 0$. Therefore, τ as a parameter of $\delta(t - \tau)$, is not directly learnable in this manner. To solve this problem, we approximate the δ function with a smoother function. Below, we show that the sinc function is an appropriate approximation of the dirac-delta function for generating continuous emotion curves; sinc function generates smooth curves that are consistent with the slow-moving ground-truth curves of emotions.

Another issue is that the mapping function, $\mathcal{F}_{pred}$, may generate a signal that contains high-frequency components, which is not desirable in continuous emotion recognition because human annotations are smooth and slow-moving. In our previous paper, we showed that incorporating a temporal smoothing technique into the network architecture can improve the performance [32]. We propose to apply a low-pass filter, $h_{lp}(t)$, to the signal generated by $\mathcal{F}_{pred}$, i.e.,

$$y(t) = (\mathcal{F}_{pred}(X(t)) * h_{lp}(t)) * \delta(t - \tau), \qquad (4)$$

which is equivalent to

$$y(t) = \mathcal{F}_{pred}(X(t)) * h_{lp}(t - \tau). \qquad (5)$$

Accordingly, we can use a time-shifted low-pass filter instead of a dirac-delta function to compensate for reaction lag and also remove the unsatisfactory high-frequency components from the output. An ideal low-pass filter, $h_{lp}(t)$, is the $sinc$ function, which can be expressed by:

$$h_{lp}(t) = 2f_c \, sinc(2f_c t),$$
$$sinc(t) = \begin{cases} \frac{\sin(\pi t)}{\pi t} & t \neq 0 \\ 1 & t = 0 \end{cases}, \qquad (6)$$

where f_c is the cutoff frequency (also known as bandwidth), which is defined as the maximum frequency that the sinc filter does not attenuate.

The cutoff frequency of the sinc filter, f_c, must be higher than the maximum frequency of the ground-truth signal, f_g, (i.e., $f_c \geqslant f_g$). Otherwise, the sinc filter cannot pass all frequencies of the predicted emotional labels and the sinc output will be smoother than the actual ground-truth labels.

For a real discrete-time signal, the frequency components ranges from 0 to $\frac{f_s}{2}$, where f_s is the sampling frequency. In

this case, the sinc filter with $\frac{f_s}{2}$ cutoff frequency passes all frequencies without attenuating them. Therefore, sinc filter with $\frac{f_s}{2}$ cutoff frequency can be used to apply a delay to any real signal sampled at f_s.

The sinc filter expressed by equation (6) has infinite number of coefficients and therefore it is not implementable in practice. In order to approximate it, a windowed-sinc filter is commonly used instead of the ideal low-pass filter [69], [70]. Equation (7) expresses the input-output relationship after applying a window $h_w(t)$ to the sinc filter:

$$y(t) = \mathcal{F}_{pred}(X(t)) * \left(2f_c \, sinc(2f_c(t - \tau)) \, h_w(t) \right) \quad (7)$$

Applying $h_w(t)$ causes distortion to the ideal frequency response of the sinc filter. In our initial experiments, we noticed that the type of the window does not significantly change the final predictions; therefore, we employ a simple rectangular window in our experiments. Equation (7) expresses our convolutional layer in the continuous time domain. To implement it, we must discretize Equation (7) using the sampling frequency of f_s (25Hz for RECOLA and 10Hz for SEWA), which leads to the following convolutional kernel for our delayed sinc layer:

$$h_{sinc}[n; \tau] = (2f_c/f_s) \, sinc[2f_c(n/f_s - \tau)] \, h_w[n]. \quad (8)$$

In summary, the delayed sinc layer is a convolutional layer that uses a special kernel. The shape of the kernel is limited to a time-shifted sinc, expressed by equation (8). Delayed sinc layer has 3 parameters:

1) Delay parameter, τ: Delayed sinc layer introduces a delay of τ seconds to its input. τ is the only parameter of the delayed sinc layer that has to be trained. In our experiments, we initialize τ randomly using a uniform distribution between 0 and 20 seconds.

2) Bandwidth of the sinc kernel, f_c: It is a constant parameter that must be higher than the bandwidth of the ground-truth signals. We set this parameter to $\frac{f_s}{2}$ in our experiments, where f_s is the sampling frequency of the ground-truth signals (i.e., 25Hz for the RECOLA dataset and 10Hz the SEWA dataset)

3) Windowing function, $h_w[n]$: We use a long rectangular window with the length of 44 seconds to be sure that the window does not remove the main beam of the sinc function even after applying the longest initial delay (i.e., 20 seconds). Additionally, this window results in a network with 44 seconds receptive field which is consistent with the effective receptive field found in [32].

6 MULTI-DELAY SINC NETWORK

In the previous section, we introduced a convolutional kernel that can compensate for a time-invariant delay. Delays introduced by human annotators, however, are not necessarily constant in time; annotators may introduce different delays to different regions in the input. For example, it is easier to identify emotions for laughter parts of speech [71], [72] and therefore annotators may be able to identify them faster. This section introduces a new network architecture, the *multi-delay sinc (MDS)* network, that utilizes multiple delayed sinc layers to deal with time-variant reaction delays.
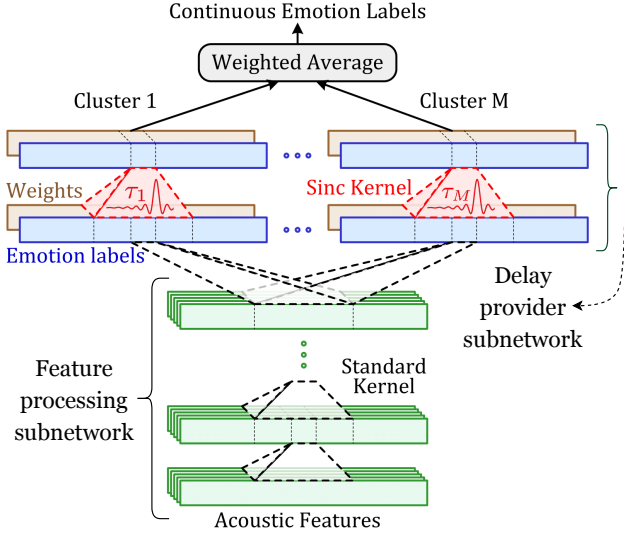
Fig. 2. A visualization of our multi-delay sinc network with M clusters. τ_m is the delay considered for the m-th component. Standard and sinc kernels are shown in black and red colors, respectively. In this figure, we show the structure of the first and the last clusters.

We first describe how multiple delayed sinc layers are used. Each delayed sinc layer is applied to a different region of the acoustic space, formed by generating fuzzy clusters. We then describe a network architecture that can integrate the prediction from multiple layers.

6.1 Acoustic Clustering

Figure 2 shows the architecture of the MDS network with M clusters. The main idea of the MDS network is to categorize speech regions into a number of fuzzy clusters such that all samples associated with a cluster require the same delay to be synchronized with the ground-truth labels. The system is trained in an end-to-end manner and fuzzy membership functions of clusters are implicitly learned. The number of clusters, M, is a parameter that must be tuned. More precisely, the MDS network defines three components for each cluster m:

1) τ_m: a learnable delay which is a single parameter for each cluster. τ_m will be trained along with other parameters of the network.

2) $\mathcal{F}_m$: emotion recognition, a mapping that predicts emotion labels for the m-th cluster based on input features, X. $\mathcal{F}_m[n; X]$ denotes the n-th sample of the signal generated by the mapping $\mathcal{F}_m$. We employ a standard multi-layer convolutional neural network to generate $\mathcal{F}_m[n; X]$ from X.

3) w_m: a mapping that generates a weight signal for the m-th cluster using input features, X. $w_m[n; X]$ denotes the n-th sample of the signal generated by the w_m mapping. $w_m[n; X]$ quantifies the importance of incorporating $\mathcal{F}_m[n; X]$ into the final predictions $y[n]$. A standard convolutional neural network is employed to define the mapping w_m.

6.2 Network Architecture

The proposed MDS network takes MFB features as input and passes them through a stack of three subnetworks: (1) feature processing, (2) delay provider, and (3) averaging subnetworks. The subnetworks are shown in Figure 2.

6.2.1 Feature processing subnetwork

The feature processing subnetwork takes acoustic features and generates emotion labels, $\mathcal{F}_m[n; X]$, and weight signals, $w_m[n; X]$ for all clusters. The MDS network uses a shared multilayer convolutional network to simultaneously generate all the emotion labels ($\mathcal{F}_m[n; X]$) and weight signals ($w_m[n; X]$). Our initial experiments showed that using a separate network to generate emotion labels and weight signals does not improve the results. The output of this network is M label signals and M weight signals, where M is the number of clusters. Each label and weight signal is a one-dimensional signal with the same length of the final emotion predictions.

6.2.2 Delay provider subnetwork

This subnetwork applies a cluster-specific delay, τ_m, to both labels ($\mathcal{F}_m[n; X]$) and weights ($w_m[n; X]$) using a delayed sinc kernel. Suppose $\vec{\mathcal{F}}_m[n; X, \tau_m]$ and $\vec{w}_m[n; X, \tau_m]$ are the delayed labels and weights predicted for m-th cluster. Then,

$$\vec{\mathcal{F}}_m[n; X, \tau_m] = \mathcal{F}_m[n; X] * h_{sinc}[n; \tau_m], \qquad (9)$$

$$\vec{w}_m[n; X, \tau_m] = w_m[n; X] * h_{sinc}[n; \tau_m], \qquad (10)$$

where $h_{sinc}[n; \tau_m]$ is the windowed sinc kernel expressed by Equation (8). This subnetwork generates a series of predictions that are hypothesized to be more closely aligned with the input features. This subnetwork has M parameters $\{\tau_m \ \forall \ m \in \{1...M\}\}$ that have to be trained.

6.2.3 Averaging subnetwork

The previous sections explained how feature processing and delay provider subnetworks generate weights $\vec{w}_m[n; X, \tau_m]$ and emotion labels $\vec{\mathcal{F}}_m[n; X, \tau_m]$ for the m-th cluster. The goal of the averaging sub-network is to generate final emotion labels $y[n]$ by combining cluster-dependent labels through cluster weights. One straightforward approach is to use the emotions predicted by the most likely cluster (the cluster with maximum weight); i.e.,

$$\hat{m}[n] = \arg\max_m \vec{w}_m[n; X, \tau_m], \qquad (11)$$

$$y[n] = \vec{\mathcal{F}}_{\hat{m}[n]}[n; X, \tau_{\hat{m}[n]}], \qquad (12)$$

where $\hat{m}[n]$ is the index of the most likely cluster (cluster with maximum weight) at time n. However, there are two problems with this approach: (1) the assumption that a part of the signal is associated with a single fixed delay is restrictive; and (2) $\hat{m}[n]$ may change at the middle of a recording and as a result predicted labels may experience a sudden change in a recording which is not consistent with the nature of the emotion labels. To deal with this problem, we propose to use a soft-max instead of the standard max. By using soft-max all clusters will contribute in generating the final emotion labels, and therefore it is less likely to observe sudden changes in the final emotion labels.

In the proposed network, the final continuous emotion label, $y[n]$, is obtained by taking a weighted average of the cluster-specific predictions. The parameters used to weight the predictions are derived from the weights described in Section 6.1. We convert these weights in the previous section to a probability distribution by passing them through a time-distributed softmax layer, which applies a softmax function at each time-step. Let $\tilde{w}_m[n; X, \tau_m]$ denote the output of the softmax layer, then,

$$\tilde{w}_m[n; X, \tau_m] = \frac{\exp(\vec{w}_m[n; X, \tau_m])}{\sum_{m'=1}^{M} \exp(\vec{w}_{m'}[n; X, \tau_{m'}])}. \quad (13)$$

The averaging network uses the cluster probabilities obtained through equation (13) to calculate the final predictions, $y[n]$:

$$y[n] = \sum_{m=1}^{M} \tilde{w}_m[n; X, \tau_m] \vec{\mathcal{F}}_m[n; X, \tau_m]. \quad (14)$$

The averaging subnetwork does not have any parameters to be trained in the training phase.

We train all these subnetworks in an end-to-end manner using the back-propagation algorithm. The result is the MDS network that automatically assigns features to clusters, applies delay to each cluster, and aggregates the result.

Predicted emotion labels for each cluster ($\vec{\mathcal{F}}_m[n; X, \tau_m]$) are band-limited signals with the maximum frequency of f_c (i.e., cut-off frequency of the sinc filter); however, when we combine them using time-varying weights, the generated continuous emotion labels, $y[n]$, may have higher frequency components. Using multiple clusters enable us to provide more complex delay components, but it has a disadvantage too; it may generate high frequency components in the output, which is not consistent with the slow moving nature of the continuous emotion labels.

7 EXPERIMENTS

7.1 Experimental Setup

We build our models using the TensorFlow numerical computation library [73]. We train all models by optimizing the CCC metric through the Adam optimizer [68], [74]. Each network is trained for 300 epochs and the best epoch is selected during validation. To reduce the effect of random initialization, we train each network three times and select the best performing network based on the validation CCC. We use the AVEC and leave-one-speaker-out schemes, explained in Section 3.2, to calculate validation performance, tune hyper-parameters, and evaluate networks.

7.1.1 Baseline System

We implement the RECOLA state-of-the-art audio system, the downsampling/upsampling convolutional network, as the baseline method for comparison, introduced in our prior work [32]. The network first encodes the input signal into a low-resolution signal through a stack of convolution max-pooling layers and then reconstructs the output through a stack of upsampling layers. We exploit transposed convolution layers (deconvolution layers) to upsample the encoded representations and generate the output annotations. We

apply the $tanh$ function after each layer (except the final layer which is a linear layer).

We train our downsampling/upsampling network by fixing the learning rate, down-sampling ratio, and number of downsampling layers to 0.0001, 2 and 7, respectively. We also cross-validate number of kernels (32, 64, 128), length of kernels (3, 4, 5) and L2 regularization weight (0.0, 0.02, 0.04) based on the validation CCC. We selected these values based on the validation CCC results that we obtained in our initial experiments.

7.1.2 MDS System

We implement the proposed MDS system as described in Section 6. We apply $tanh$ nonlinearity after all standard convolutional kernels, except the ones that generate cluster weights and labels. We cannot apply any nonlinearity function after delayed sinc layers because sinc kernels have been specifically designed to generate frequency components of the ground-truth labels and applying nonlinearities will change their frequency response. We train the MDS network by fixing the learning rate to 0.001 and cross-validating the number of the standard kernels (16, 32), length of the standard kernels (4, 8, 16), number of the convolution layers (3, 5) and L2 regularization weight (0.0, 0.025) based on the validation performance.

We conduct an experiment to select a good value for f_c, the cutoff frequency of the sinc kernels. We apply a windowed-sinc filter with different f_c to all training labels and compare the resulting labels with the original ones using the CCC metric (Figure 3). The results obtained for RECOLA and SEWA are very similar, showing that the ground-truth labels in both datasets share similar frequency characteristics; for example, selecting f_c greater than 0.5Hz results in a less than 1 percent reduction in CCC on both datasets. Therefore, a number higher than 0.5Hz is a good choice for f_c. We set f_c to $f_s/2$ in our experiments, where f_s is the sampling rate of the emotion labels (25Hz for RECOLA and 10Hz for SEWA).

We apply 32 delayed sinc kernels of length of 44 seconds. We discuss the effects of kernel length on performance in Section 7.3. We initialize the delay parameters τ_m of the sinc kernels randomly, between 0 and 20 seconds. Our initial experiments show that the uniform distribution is better that the Gaussian distribution for initializing the delay values. We find that when we initialize the delay values with negative numbers, they tend to converge to positive numbers. This supports the claim that the delay values approximate the reaction delays of annotators.

7.2 Results

We present the results obtained for the RECOLA and the SEWA datasets separately in this section.

7.2.1 RECOLA

Table 1 (RECOLA) reports the leave-one-speaker-out CCCs calculated for both MDS and downsampling/upsampling networks. The results show that our system is significantly better than the downsampling/upsampling network for both arousal and valence. Our arousal predictions exhibit 0.02±0.024 improvement (pairwise t-test, p-
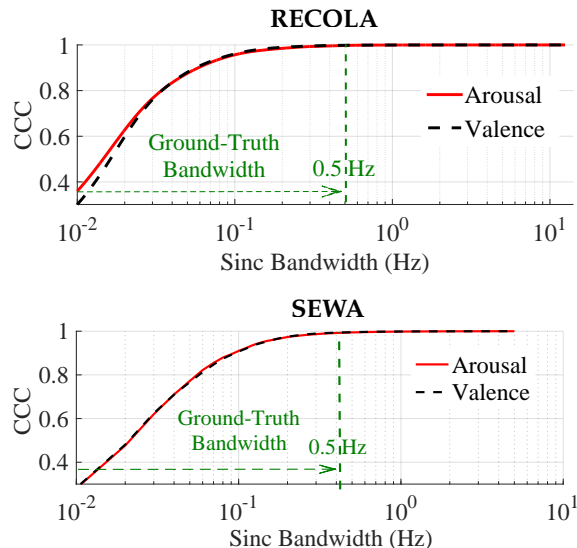
Fig. 3. Performance reduction caused by applying a windowed-sinc filter to the ground-truth labels on both RECOLA and SEWA datasets. *0.5Hz* can be considered as the bandwidth of the emotion labels.

value=0.04). Our valence predictions show the improvement of 0.056±0.07 (pairwise t-test, p-value=0.047).

We also report the test CCC results calculated according to the AVEC evaluation scheme explained in section 3.2. We rank all trained systems in the previous experiment based on their development CCC. We then select the best performing MDS and downsampling/upsampling networks based on the development CCC. We use the selected systems to generate the emotion labels for test utterances. For both the arousal and valence prediction tasks, the best MDS network contains 5 layers with 16 filters in each layer and 8 filter coefficients for each filter. The best downsampling/upsampling network for arousal has 7 layers, 32 filters in each layer and 3 filter coefficients for each filter. It also uses L2 regularization factor of 0.02. The best for valence has 7 layers, 128 filters, 3 filter coefficients and 0.04 L2 regularization factor.

Table 2 (RECOLA) summarizes the development and test results calculated according to the AVEC evaluation scheme. For the arousal prediction, all results are slightly in favor of the proposed system. For the valence prediction, our proposed system improves CCC values but it cannot improve RMSE values. We hypothesize that it is because we train the networks to maximize the CCC metric, which does not necessarily lead to a good RMSE estimator.

7.2.2 SEWA

Table 1 (SEWA) compares the CCC result of MDS with downsampling/upsampling networks for each subject. Similar to the RECOLA dataset, our system outperforms the downsampling/upsampling network for most of the subjects (10/14 subjects for predicting arousal and 9/14 subjects for predicting valence). Our system exhibits an improvement of 0.059 ± 0.13 on arousal prediction and 0.096 ± 0.13 on valence recognition. The improvement is not significant for the arousal prediction (p-value=0.11), but it is significant for the valence prediction (p-value=0.02).

We also assess the performance of MDS and downsampling/upsampling networks using AVEC evaluation

TABLE 1
CCC results of downsampling/upsampling (down/up) and MDS networks for each subject in RECOLA and SEWA datasets. leave-one-speaker-out evaluation scheme is used to calculate the CCC results for each subject. MDS network fails to improve the downsampling/upsampling network for the shaded subjects.

RECOLA

Sub	Arousal		Valence	
	Down/up	MDS-net	Down/up	MDS-net
1	.812	**.820**	.429	**.494**
2	.922	**.938**	.365	**.440**
3	.800	**.835**	.370	**.493**
4	.755	**.787**	.135	**.163**
5	.812	**.875**	.399	**.485**
6	.821	**.858**	.445	**.463**
7	**.846**	.844	.265	**.447**
8	**.637**	.615	**.706**	.657
9	.744	**.762**	**.658**	.634
$\mu \pm \sigma$	.794±.08	**.814±.09**	.419±.18	**.475±.14**

SEWA

Sub	Arousal		Valence	
	Down/up	MDS-net	Down/up	MDS-net
1	.368	**.488**	.471	**.671**
2	.325	**.400**	**.228**	.218
3	**.174**	-.036	**.184**	.043
4	.372	**.579**	.455	**.697**
5	**.512**	.510	**.480**	.454
6	.336	**.646**	.373	**.662**
7	.616	**.682**	.532	**.601**
8	**.566**	.529	.350	**.601**
9	.031	**.033**	**.062**	.037
10	.474	**.525**	.286	**.472**
11	.297	**.373**	.321	**.501**
12	**.241**	.153	**.396**	.351
13	.516	**.615**	.170	**.327**
14	.412	**.567**	.392	**.406**
$\mu \pm \sigma$	.374±.16	**.433±.23**	.336±.14	**.432±.22**

TABLE 2
Comparing down/up and MDS networks on RECOLA and SEWA datasets. AVEC evaluation scheme is used for this comparison.

RECOLA		Development		Test	
		Down/up	MDS-net	Down/up	MDS-net
CCC	Arousal	.865	**.873**	.680	**.688**
	Valence	.574	**.591**	.472	**.492**
RMSE	Arousal	.098	**.097**	.141	**.136**
	Valence	**.105**	.119	**.116**	.126

SEWA		Development		Test	
		Down/up	MDS-net	Down/up	MDS-net
CCC	Arousal	.458	**.530**	.317	**.412**
	Valence	.485	**.542**	.287	**.379**
RMSE	Arousal	.139	**.135**	.130	**.124**
	Valence	.157	**.138**	.178	**.130**

TABLE 3
Comparing emotion recognition systems on the test set of SEWA.

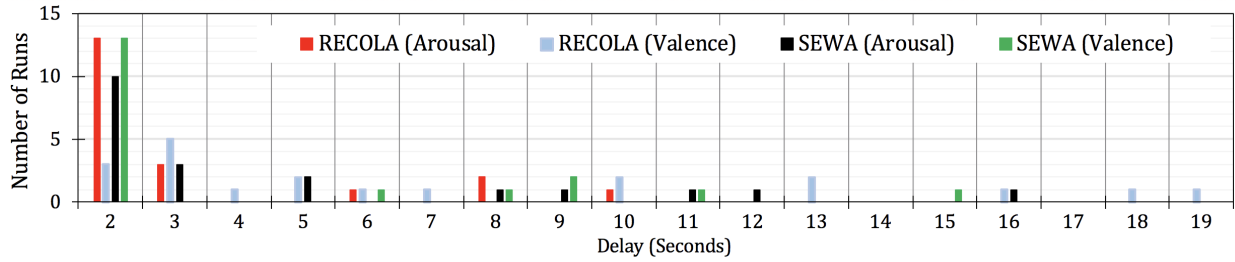| Methods	Arousal		Valence	
	RMSE	CCC	RMSE	CCC
Down/up	.130	.317	.178	.287
eGeMAPS-GMR [75]	–	.344	–	.346
MDS-net	.124	.412	.130	.379
IS10-LSTM [37]	**.100**	.422	**.112**	.405
IS10-LSTM + MDS-net	**.101**	**.458**	**.114**	**.421**

Fig. 4. Distribution of the delays trained through different runs of the MDS network with one cluster. This network learns one delay to compensate for the reaction lags.

TABLE 4
Comparing space complexity of the best networks trained on both SEWA and RECOLA. Two numbers are reported in each cell: number of training parameters and applied L2 regularization factor. For example, 56K(.02) shows the best network has 56,000 parameters and is trained with an L2 factor of 0.02.

| | RECOLA | | SEWA | |
Networks	Arousal	Valence	Arousal	Valence
Down/up	56K (.02)	703K (.04)	290K (.04)	1,415K (.04)
MDS-net	**37K (0)**	**37K (0)**	**119K (.025)**	**57K (.025)**

scheme, explained in section 3.2. We train our networks with different hyper-parameters explained in section 7.1 on the training partition of SEWA. We then select the best performing network based on the development CCC.

The best MDS arousal predictor contains 3 layers with 32 filters of length 4. This MDS network is trained with L2 regularization factor of 0.025. The best MDS valence predictor differs in only one parameter: it has 16 filters in each layer. The best downsampling/upsampling network for arousal prediction contains 64 convolutional filters of length 3 in each layer with L2 factor of 0.04. For valence prediction, it has 128 convolutional filters of length 5 with L2 factor of 0.04.

Table 2 shows the development and test results of the selected networks. This table confirms that our network outperforms the downsampling/upsampling network for both arousal and valence prediction tasks on both development and test partitions of the database using both comparison metrics (i.e., RMSE and CCC).

However, the downsampling/upsampling network is not a state-of-the-art method on the SEWA dataset. We also compare the performance of the proposed MDS network with two other emotion recognition systems: eGeMAPS-GMR [75] and IS10-LSTM [37]. Dang et al. [75] used eGeMAPS features [51] with Gaussian mixture regression (GMR) [76] to recognize emotion labels in SEWA. Chen et al. [37] used the IS10 [54] feature set to train a multi-task LSTM network that predicts both arousal and valence labels simultaneously. We note that both systems outperform the downsampling/upsampling network and that while MDS outperforms the eGeMAPS-GMR system, it is outperformed by the IS10-LSTM system (Table 3).

We analyzed the predictions generated by MDS and IS10-LSTM networks and observed that although both networks are accurate, the generated predictions are often not highly correlated[4] (Their CCC for arousal and valence are

4. Many thanks to the authors of the IS10-LSTM [37] paper for sending us predictions of their systems.

0.282 and 0.362, respectively; this correlation is calculated on the development set). This suggests that the two approaches may be considering different aspects of the input signal when predicting the labels. We hypothesized that we could improve the predictions by fusing the results of the two systems. We performed the fusion by taking the average of the predictions. We refer to this system as the "IS10-LSTM + MDS-net" system and find that this system considerably improves the CCC of both "IS10-LSTM" and "MDS-net" systems while preserving the RMSE of the "IS10-LSTM" network (Table 3).

Table 4 compares the memory complexity of our best downsampling/upsampling and MDS networks. The number of parameters and the L2 regularization factor of each network are shown in the table. According to the table, the downsampling/upsampling structure requires a larger network with higher regularization factor to predict emotion labels of both RECOLA and SEWA. This high regularization factor is crucial for the large downsampling/upsampling networks to reduce their over-fitting problem. The table confirms that the MDS network outperforms the downsampling/upsampling network with fewer parameters.

7.3 Structural Analysis of MDS Network

In this section, we analyze various aspects of the MDS network. We try to answer the following questions:

- Is delayed sinc layer able to compensate for reaction lags and synchronize speech with emotion labels? (Sec. 7.3.1)
- What is the effective range for the bandwidth parameter in delayed sinc layers? (Sec. 7.3.2)
- How many clusters are required to train a robust emotion recognition system? (Sec. 7.3.3)
- What is the maximum delay component that can affect continuous emotion recognition? (Sec. 7.3.4)
- Do the reaction delays change with different acoustic events? (Sec.7.3.5)

7.3.1 Learning delay through delayed sinc layer

In Section 4, we found an estimate of the reaction delay using a brute-force algorithm. The algorithm trains a separate network for any candidate value of delay and selects the delay that leads to the best CCC result. This approach is resource intensive and also it is not suitable for training multiple delays. We demonstrate that the delayed sinc layer can learn comparable delays in a less resource intensive manner using the back-propagation algorithm.

We train our MDS network to learn a single delay by fixing the number of clusters to one (one delay parameter).

We will compare the learned delay to the delay found in Section 4. We train twenty networks with random initializations and study the distribution of the learned delay values. We expect that the final delay values will be similar to the values we obtained in Section 4. We set all parameters, except the number of clusters, of the networks (e.g., number of layers, length of filters, etc.) to the parameters of the best networks introduced in Section 7.2.

Figure 4 shows the distribution of the trained delays for both arousal and valence predictions on both RECOLA and SEWA datasets. Although we initialize the delays randomly between 0 to 20 seconds, most of the delays tend to converge to a number between 1.5 and 3.5 seconds. This interval agrees with the optimal delays obtained through exhaustive search in Section 4.

It is important to quantify the likelihood of finding an effective reaction delay (a delay between 1.5 and 3.5 seconds) using delayed sinc kernel. This likelihood measures the ability of the proposed delayed sinc kernel in finding and compensating for the reaction delays. According to Figure 4 this likelihood is different for different datasets (RECOLA, SEWA) and different tasks (arousal, valence). We quantify the likelihood by calculating the percentage of runs in which the trained delay parameter is a number between 1.5 and 3.5 seconds. This percentage is equal to 80%, 45%, 65% and 65% for RECOLA-arousal, RECOLA-valence, SEWA-arousal and SEWA-valence, respectively.

The results of this section confirm that the delayed sinc layer is able to learn and compensate the reaction delays in most runs. However, in some runs the sinc layer does not converge to the optimal solution of the exhaustive search. It is because our optimization function (CCC) has multiple local optima and gradient-based optimization algorithms may get stuck in the local optima.

7.3.2 Sinc bandwidth

The sinc bandwidth parameter can be learned during the training process, but we set it to a constant value (1Hz) throughout the experiments, because our initial experiments showed that learning this parameter does not improve the final CCC. We now explore the relationship between the final performance and the value of this parameter. To this end, we train our best MDS network explained in the previous section with different sinc bandwidths ranging exponentially from 2^{-7} to the maximum frequency component (12.5Hz for RECOLA and 5Hz for SEWA). We train each network 10 times with different random initializations and report the average of the leave-one-speaker-out values in Figure 5. As can be seen in the figure, any bandwidth higher than 0.125Hz for arousal and 0.5Hz for valence results in good performance. According to Section 5.1 and Figure 3, 0.5Hz is the bandwidth of the continuous emotion labels. Therefore, we do not need to learn the sinc bandwidths during the training process and we can just select a frequency higher than the output bandwidth (i.e., 0.5Hz). Also, Figure 5 shows that Valence prediction is more sensitive to attenuating frequencies between 0.125Hz and 0.5Hz.

7.3.3 Number of the clusters

Our system divides the acoustic space into several fuzzy clusters and applies different delays to each cluster. In this
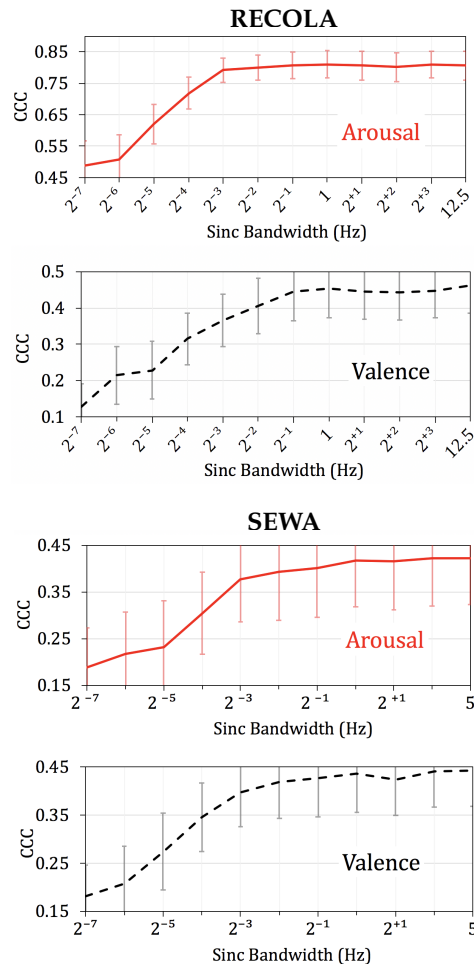

Fig. 5. Increasing bandwidth of the sinc kernels up to 0.125Hz for arousal and 0.5 Hz for valence improves the leave-one-speaker-out performance. Increasing more does not significantly change the CCC.

section, we investigate the utility of this approach. We train the best MDS network explained in Section 7.2 for different number of clusters ranging exponentially from 1 to 128 and compare their leave-one-speaker-out CCC. Figure 6 shows the results obtained in this experiment for both RECOLA and SEWA datasets. According to the results, using one cluster is not enough and the MDS network needs at least 8 clusters for arousal prediction and 16 clusters for valence prediction to provide a high performance system that is comparable to the best performing system. The results also show that more clusters are needed to predict valence, compared to arousal.

7.3.4 Maximum delay parameter

In this section, our goal is to find the maximum delay that can assist continuous emotion recognition. To this end, we train different networks with different maximum delays, τ_{max}, and report the development CCCs of the networks with respect to τ_{max}. More specifically, we set the maximum initial delay to τ_{max}, the length of delayed sinc kernels to $2\tau_{max}$, and all other hyper-parameters to the best values reported in Section 7.2. We train and evaluate the networks on SEWA using the AVEC evaluation scheme explained in
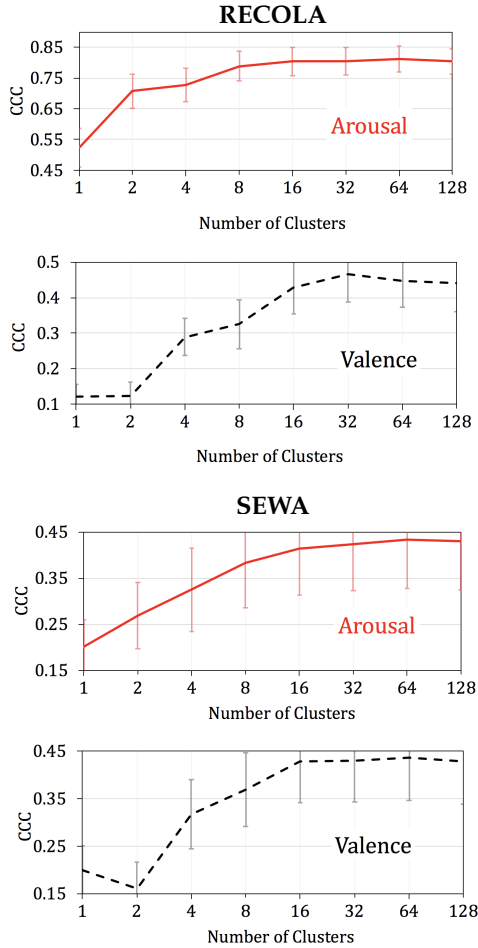
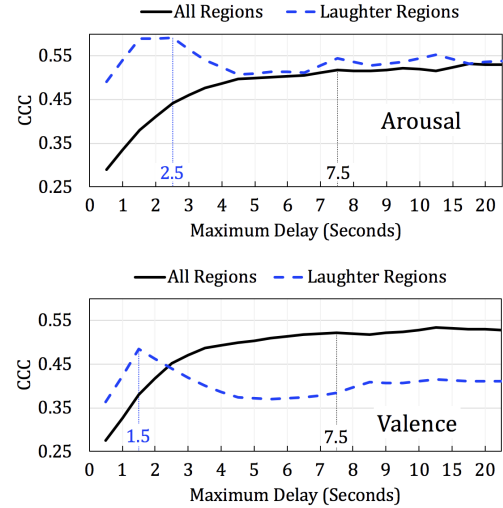Fig. 6. CCC results for different number of clusters.

Fig. 7. CCC results of the proposed network with different maximum delays on the SEWA dataset. AVEC evaluation scheme, explained in Section 3.2, is used to calculate the CCC values of the laughter parts (dashed blue curves) and all parts (solid black curves) of the development set.

Section 3.2. We train each network ten times and report the average CCC to reduce the effect of random initialization.

Figure 7 (solid black curves) shows the results of this experiment for both arousal and valence prediction. According to this figure, CCC values improve consistently by increasing the maximum delay up to 7.5 seconds. Therefore 7.5 seconds can be considered as the maximum delay that can assist emotion recognition on SEWA.

7.3.5 Effect of acoustic events on delays

The proposed MDS network compensates for annotators' delays by clustering acoustic space into several fuzzy clusters and by introducing different delays to different clusters; therefore, the network assumes that the delays depend on acoustic clusters. In this section, we explore this assumption, asking if the annotators' delays change with different regions in the acoustic space.

We study the reaction delays for laughter regions of signal and compare them with the delays of other parts of the signal. We select laughter because it is highly likely that laughter requires smaller reaction delays compared to speech. Many studies discussed acoustic characteristics of laughter and concluded that there are several distinguishable features in laughter; for example, laughter has longer unvoiced portions than voiced portions [71], [72]. These

features can facilitate identifying laughter and can reduce the annotators' reaction times.

In order to find the delays of the laughter parts, we repeat the experiment reported in the previous section, but with the difference that we calculate the CCC results just for laughter parts of the signals. Figure 7 (dashed blue curves) demonstrates these CCC results with respect to the maximum delay parameter. Increasing the maximum delay parameter more than 2.5 seconds for arousal and 1.5 seconds for valence considerably reduces the CCC values. It shows that predicting emotion labels of the laughter regions can be done by applying smaller delay components compared to other parts of a speech signal. We hypothesize that it is because human annotators have smaller reaction delays in identifying emotion labels of the laughter regions and therefore reaction delays depend on acoustic variability.

8 DISCUSSION

Annotators' reaction delays make continuous emotion recognition challenging because the delays introduce non-additive (convolutive) noise components to the emotional labels and create asynchronous ground-truth signals. Compensating for these delays is not straightforward as the delays depend on different factors such as the age of the annotators [77], concentration level of the annotators [78], and affective behaviours [23].

The proposed network models the reaction delays through functions of affective behaviours captured by acoustic clusters. The network categorizes acoustic features into several fuzzy clusters (membership degrees can be any number between zero and one) and applies different delays for different clusters. Since acoustic clusters vary over time, the modeled reaction delay is also time-varying.

The results reported in Table 1 demonstrate the superiority of our network over the downsampling/upsampling network [32]. On the RECOLA dataset, our system is better for most of the subjects (7 out of 9 subjects), but cannot

improve the performance on subject 8. We studied the recorded audio files and we noticed that for all files in the development set, the interviewee's voice dominates the leaked interviewer's voice, but for subject 8 the power of the interviewer's voice is comparable to the power of the interviewee's voice. This considerable leakage may be a reason that our system behaves differently for this subject.

The leave-one-speaker-out experiment (Table 1) shows that the CCC results of arousal recognition are very different for RECOLA and SEWA datasets (CCC result is around 0.8 on RECOLA, but around 0.4 on SEWA). In addition, the MDS network performs similarly in predicting arousal (0.433 CCC) and valence (0.432 CCC) on SEWA, but not on RECOLA; predicting arousal is easier on RECOLA. One important reason for these inconsistencies is that RECOLA and SEWA use different methods for collecting emotion labels. The differences between them were discussed in section 3.1.3.

9 CONCLUSION

This paper introduces a new method to align acoustic features and continuous emotion labels using the delayed sinc layer. This layer is able to introduce a learnable delay to its input. Our experiments show that the delayed sinc layer can successfully align two signals by introducing a single (uniform) delay to one of them. However, a uniform delay is not enough for aligning speech and emotion annotations, because reaction delays of annotators may vary with affective behaviours. To deal with this issue, we combine multiple delayed sinc layers into a network architecture. The network categorizes features into a number of fuzzy clusters (it is fuzzy because each acoustic feature can belong to more than one cluster), and then learns different delays for different acoustic clusters. Our experiments show that: (1) the sinc filter with a cutoff frequency higher than the bandwidth of the ground-truth signal can be used to deal with the misalignment problem; (2) the proposed network significantly outperforms the downsampling/upsampling network; (3) predicting valence requires more clusters compared to predicting arousal. (4) laughter requires smaller delay components compared to other regions of speech.

We used a sinc filter to approximate the dirac-delta function. However, other functions, such as Gaussian and triangular, can also be employed instead of the sinc kernel. Future work will explore the effect of using different types of kernels that can approximate the dirac-delta function. Additionally, in this paper, we focused on the speech modality to predict continuous emotion annotations, while the proposed multi-delay sinc network is a reasonable modeling technique for other input modalities too. Another future plan is to evaluate the performance of the proposed network over other physiological and behavioral modalities such as: video [79], [80], body language [81], [82] and EEG [83], [84].

ACKNOWLEDGMENT

REFERENCES

[1] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[2] J. Gideon, B. Zhang, Z. Aldeneh, Y. Kim, S. Khorram, D. Le, and E. M. Provost, "Wild wild emotion: a multimodal ensemble approach," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 501–505.

[3] D. Cordaro, R. Sun, D. Keltner, S. Kamble, N. Huddar, and G. McNeil, "Universals and cultural variations in 22 emotional expressions across five cultures." *Emotion (Washington, DC)*, vol. 18, no. 1, pp. 75–93, 2018.

[4] B. W. Schuller, "Speech emotion recognition: two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.

[5] A. Mencattini, E. Martinelli, F. Ringeval, B. Schuller, and C. Di Natale, "Continuous estimation of emotions in speech by dynamic cooperative speaker models," *IEEE transactions on affective computing*, vol. 8, no. 3, pp. 314–327, 2017.

[6] F. Ringeval, B. Schuller, M. Valstar, S. Jaiswal, E. Marchi, D. Lalanne, R. Cowie, and M. Pantic, "Av+ec 2015: the first affect recognition challenge bridging across audio, video, and physiological data," in *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2015, pp. 3–8.

[7] Z. Aldeneh, S. Khorram, D. Dimitriadis, and E. M. Provost, "Pooling acoustic and lexical features for the prediction of valence," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 68–72.

[8] G. Keren, T. Kirschstein, E. Marchi, F. Ringeval, and B. Schuller, "End-to-end learning for dimensional emotion recognition from physiological signals," in *Multimedia and Expo (ICME), 2017 IEEE International Conference on*. IEEE, 2017, pp. 985–990.

[9] M. Abdelwahab and C. Busso, "Study of dense network approaches for speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018), Calgary, AB, Canada*, 2018.

[10] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," *Proc. Interspeech 2018*, pp. 3698–3702, 2018.

[11] B. Zhang, S. Khorram, and E. M. Provost, "Exploiting acoustic and lexical properties of phonemes to recognize valence from speech," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5871–5875.

[12] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[13] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, and M. Pantic, "Summary for avec 2017: real-life depression and affect challenge and workshop," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 1963–1964.

[14] J. Chang and S. Scherer, "Learning representations of emotional speech with deep convolutional generative adversarial networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2746–2750.

[15] C. Busso, M. Bulut, S. Narayanan, J. Gratch, and S. Marsella, "Toward effective automatic recognition systems of emotion in speech," *Social emotions in nature and artifact: emotions in human and human-computer interaction, J. Gratch and S. Marsella, Eds*, pp. 110–127, 2013.

[16] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[17] R. Cowie and R. R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech communication*, vol. 40, no. 1-2, pp. 5–32, 2003.

[18] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," *INTERSPEECH, Stockholm, Sweden*, 2017.

[19] B. Zhang, E. M. Provost, and G. Essl, "Cross-corpus acoustic emotion recognition with multi-task learning: seeking common ground while preserving differences," *IEEE Transactions on Affective Computing*, 2017.

[20] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Prediction-based learning for continuous emotion recognition in speech," in *Acous-*

tics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017, pp. 5005–5009.

[21] J. Huang, Y. Li, J. Tao, Z. Lian, Z. Wen, M. Yang, and J. Yi, "Continuous multimodal emotion prediction based on long short term memory recurrent neural network," in Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. ACM, 2017, pp. 11–18.

[22] Z. Huang, T. Dang, N. Cummins, B. Stasak, P. Le, V. Sethu, and J. Epps, "An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction," in Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. ACM, 2015, pp. 41–48.

[23] S. Mariooryad and C. Busso, "Correcting time-continuous emotional labels by modeling the reaction lag of evaluators," IEEE Transactions on Affective Computing, vol. 6, no. 2, pp. 97–108, 2015.

[24] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in Proceedings of the 14th ACM international conference on Multimodal interaction. ACM, 2012, pp. 501–508.

[25] E. Böring, "A history of experimental psychology," New York: Appleton-Century, 1950.

[26] J. D. Mollon and A. J. Perkins, "Errors of judgement at greenwich in 1796." Nature, 1996.

[27] S. Nicolas, "On the speed of different senses and nerve transmission by hirsch (1862)," Psychological Research, vol. 59, no. 4, pp. 261–268, 1997.

[28] S. Mariooryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on. IEEE, 2013, pp. 85–90.

[29] M. A. Nicolaou, H. Gunes, and M. Pantic, "Automatic segmentation of spontaneous data using dimensional labels from multiple coders," in Proc. of LREC Int. Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality. Citeseer, 2010, pp. 43–48.

[30] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016, pp. 5200–5204.

[31] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," IEEE Transactions on Affective Computing, vol. 2, no. 2, pp. 92–105, 2011.

[32] S. Khorram, Z. Aldeneh, D. Dimitriadis, M. McInnis, and E. M. Provost, "Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition," Proc. Interspeech 2017, pp. 1253–1257, 2017.

[33] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," Pattern Recognition Letters, vol. 66, pp. 22–30, 2015.

[34] D. Le, Z. Aldeneh, and E. M. Provost, "Discretized continuous speech emotion recognition with multi-task deep recurrent neural network." in INTERSPEECH, 2017, pp. 1108–1112.

[35] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in International Conference and Workshops on Automatic Face and Gesture Recognition (FG). IEEE, 2013, pp. 1–8.

[36] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: real-life depression, and affect recognition workshop and challenge," in Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. ACM, 2017, pp. 3–9.

[37] S. Chen, Q. Jin, J. Zhao, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," in Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge. ACM, 2017, pp. 19–26.

[38] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in Fifteenth annual conference of the international speech communication association, 2014.

[39] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," arXiv preprint arXiv:1511.07122, 2015.

[40] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015, pp. 1520–1528.

[41] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 12, pp. 2481–2495, 2017.

[42] F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud et al., "Avec 2018 workshop and challenge: bipolar disorder and cross-cultural affect recognition," in Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop. ACM, 2018, pp. 3–13.

[43] C. Wang, P. Lopes, T. Pun, and G. Chanel, "Towards a better gold standard: denoising and modelling continuous emotion annotations based on feature agglomeration and outlier regularisation," in Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop. ACM, 2018, pp. 73–81.

[44] K. Wataraka Gamage, T. Dang, V. Sethu, J. Epps, and E. Ambikairajah, "Speech-based continuous emotion prediction by learning perception responses related to salient events: a study based on vocal affect bursts and cross-cultural affect in avec 2018," in Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop. ACM, 2018, pp. 47–55.

[45] J. Zhao, R. Li, S. Chen, and Q. Jin, "Multi-modal multi-cultural dimensional continues emotion recognition in dyadic interactions," in Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop. ACM, 2018, pp. 65–72.

[46] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. ACM, 2015, pp. 73–80.

[47] F. Eyben, F. Weninger, F. Gross, and B. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in Proceedings of the 21st ACM international conference on Multimedia. ACM, 2013, pp. 835–838.

[48] B. Mathieu, S. Essid, T. Fillon, J. Prado, and G. Richard, "Yaafe, an easy to use and efficient audio feature extraction software." in ISMIR, 2010, pp. 441–446.

[49] K. Brady, Y. Gwon, P. Khorrami, E. Godoy, W. Campbell, C. Dagli, and T. S. Huang, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, 2016, pp. 97–104.

[50] F. Povolny, P. Matejka, M. Hradis, A. Popková, L. Otrusina, P. Smrz, I. Wood, C. Robin, and L. Lamel, "Multimodal emotion recognition for AVEC 2016 challenge," in Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, 2016, pp. 75–82.

[51] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," IEEE Transactions on Affective Computing, vol. 7, no. 2, pp. 190–202, 2016.

[52] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2010, pp. 2528–2535.

[53] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," arXiv preprint arXiv:1603.07285, 2016.

[54] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge," Speech Communication, vol. 53, no. 9-10, pp. 1062–1087, 2011.

[55] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: learning sound representations from unlabeled video," in Advances in Neural Information Processing Systems, 2016, pp. 892–900.

[56] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: depression, mood, and emotion recognition workshop and challenge," in Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. ACM, 2016, pp. 3–10.

[57] F. Eyben, M. Wöllmer, and B. Schuller, "Openear-introducing the munich open-source emotion and affect recognition toolkit," in Affective computing and intelligent interaction and workshops, 2009. ACII 2009. 3rd international conference on. IEEE, 2009, pp. 1–6.

[58] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," arXiv preprint arXiv:1003.4083, 2010.

[59] S. Khorram, J. Gideon, M. G. McInnis, and E. M. Provost, "Recognition of depression in bipolar disorder: Leveraging cohort and person-specific knowledge." in *INTERSPEECH*, 2016, pp. 1215–1219.

[60] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis-a unified approach to speech spectral estimation," in *Third International Conference on Spoken Language Processing*, 1994.

[61] S. Khorram, H. Sameti, F. Bahmaninezhad, S. King, and T. Drugman, "Context-dependent acoustic modeling based on hidden maximum entropy model for statistical parametric speech synthesis," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, p. 12, 2014.

[62] S. Khorram, H. Sameti, and S. King, "Soft context clustering for f0 modeling in hmm-based speech synthesis," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, p. 2, 2015.

[63] S. Khorram, F. Bahmaninezhad, and H. Sameti, "Speech synthesis based on gaussian conditional random fields," in *International Symposium on Artificial Intelligence and Signal Processing*. Springer, 2013, pp. 183–193.

[64] C. Busso, S. Lee, and S. S. Narayanan, "Using neutral speech models for emotional speech analysis." in *Interspeech*, 2007, pp. 2225–2228.

[65] S. Khorram, M. Jaiswal, J. Gideon, M. McInnis, and E.-M. Provost, "The priori emotion dataset: Linking mood to emotion detected in-the-wild," *Proc. Interspeech 2018*, pp. 1903–1907, 2018.

[66] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *workshop on automatic speech recognition and understanding (ASRU)*. IEEE, 2011.

[67] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden markov models with deep belief networks," in *workshop on automatic speech recognition and understanding (ASRU)*. IEEE, 2013, pp. 216–221.

[68] D. Kingma and J. Ba, "Adam: a method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[69] S. K. Mitra and Y. Kuo, *Digital signal processing: a computer-based approach*. McGraw-Hill Higher Education, 2006, vol. 2.

[70] S. Khorram, M. G. McInnis, and E. M. Provost, "Trainable time warping: Aligning time-series in the continuous-time domain," *arXiv preprint arXiv:1903.09245*, 2019.

[71] K. Truong and D. van Leeuwen, "Automatic detection of laughter," in *9th European Conference on Speech Communication and Technology, 4 September 2005 through 8 September 2005, Lisbon,, 485-488*, 2005.

[72] C. A. Bickley and S. Hunnicutt, "Acoustic analysis of laughter," in *Second International Conference on Spoken Language Processing*, 1992.

[73] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[74] J. Gideon, S. Khorram, Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Progressive neural networks for transfer learning in emotion recognition," *Proc. Interspeech 2017*, pp. 1098–1102, 2017.

[75] T. Dang, B. Stasak, Z. Huang, S. Jayawardena, M. Atcheson, M. Hayat, P. Le, V. Sethu, R. Goecke, and J. Epps, "Investigating word affect features and fusion of probabilistic predictions incorporating uncertainty in avec 2017," in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. ACM, 2017, pp. 27–35.

[76] A. Metallinou, A. Katsamanis, Y. Wang, and S. Narayanan, "Tracking changes in continuous emotion states using body language and prosodic cues," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2288–2291.

[77] G. Der and I. J. Deary, "Age and sex differences in reaction time in adulthood: results from the united kingdom health and lifestyle survey." *Psychology and aging*, vol. 21, no. 1, p. 62, 2006.

[78] L. Bao and E. F. Redish, "Concentration analysis: a quantitative assessment of student states," *American Journal of Physics*, vol. 69, no. S1, pp. S45–S53, 2001.

[79] H. Kaya, F. Gürpınar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image and Vision Computing*, vol. 65, pp. 66–75, 2017.

[80] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera, and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, 2017.

[81] B. Balas, A. Auen, A. Saville, and J. Schmidt, "Body emotion recognition disproportionately depends on vertical orientations during childhood." *International Journal of Behavioral Development*, vol. 42, no. 2, pp. 278–283, 2018.

[82] M. Daoudi, S. Berretti, P. Pala, Y. Delevoye, and A. Del Bimbo, "Emotion recognition by body movement representation on the manifold of symmetric positive definite matrices," in *International Conference on Image Analysis and Processing*. Springer, 2017, pp. 550–560.

[83] A. Mert and A. Akan, "Emotion recognition from eeg signals by using multivariate empirical mode decomposition," *Pattern Analysis and Applications*, vol. 21, no. 1, pp. 81–89, 2018.

[84] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from eeg," *IEEE Transactions on Affective Computing*, 2017.

Soheil Khorram is a Research Fellow in Computer Science and Engineering at the University of Michigan. He received his M.S. and Ph.D. in Computer Engineering from Sharif University of Technology, Tehran, Iran in 2008 and 2015 respectively. His M.S. and Ph.D. thesis was about "speech enhancement using subband adaptive filtering" and "improving acoustic models for statistical parametric speech synthesis". He was also a visiting student at the Center for Speech Technology Research (CSTR), University of Edinburgh, from July to December, 2013. He worked as a member of the Simple4All project under the supervision of Prof. Simon King at CSTR. He is currently working on emotion and mood recognition systems as a part of the PRIORI (Predicting Individual Outcomes for Rapid Interventions) project.

Melvin G McInnis MD, is the Thomas B and Nancy Upjohn Woodworth Professor of Bipolar Disorder and Depression and Professor of Psychiatry. He is the Director of the HC Prechter Bipolar Program and Associate Director of the Depression Center at the University of Michigan. He is a Fellow of the Royal College of Psychiatry (UK) and Fellow of the American College of Neuropsychopharmacology. Dr. McInnis trained in Canada, Iceland, England, and USA, he began a faculty position in Psychiatry at Johns Hopkins University (1993) and was recruited to the University of Michigan in 2004. His research interests include the genetics of bipolar disorder and longitudinal outcome patterns in mood disorders. He has received awards recognizing excellence in bipolar research from the National Alliance for the Mentally Ill (NAMI) and National Alliance for Research in Schizophrenia and Affective Disorders (NARSAD). He has published over 250 manuscripts related to mood disorders research, and is widely engaged in collaborative research focused on identifying biological mechanisms of disease and predictive patterns of outcomes in mental health.

Emily Mower Provost is an Associate Professor in Computer Science and Engineering at the University of Michigan. She received her B.S. in Electrical Engineering (summa cum laude and with thesis honors) from Tufts University, Boston, MA in 2004 and her M.S. and Ph.D. in Electrical Engineering from the University of Southern California (USC), Los Angeles, CA in 2007 and 2010, respectively. She is a member of Tau-Beta-Pi, Eta-Kappa-Nu, and a member of IEEE and ISCA. She has been awarded a National Science Foundation CAREER Award (2017), a National Science Foundation Graduate Research Fellowship (2004-2007), the Herbert Kunzel Engineering Fellowship from USC (2007-2008, 2010-2011), the Intel Research Fellowship (2008-2010), the Achievement Rewards For College Scientists (ARCS) Award (2009 2010), and the Oscar Stern Award for Depression Research (2015). She is a co-author on the paper, "Say Cheese vs. Smile: Reducing Speech-Related Variability for Facial Emotion Recognition," winner of Best Student Paper at ACM Multimedia, 2014, a co-author of the winner of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge, and a co-author of an honorable mention paper at ICMI 2016. Her research interests are in human-centered speech and video processing, multimodal interfaces design, and speech-based assistive technology. The goals of her research are motivated by the complexities of human emotion generation and perception.