ATTENTIVE MODALITY HOPPING MECHANISM FOR SPEECH EMOTION RECOGNITION

Seunghyun Yoon[1,3], Subhadeep Dey[2], Hwanhee Lee[1] and Kyomin Jung[1]

[1]Dept. of Electrical and Computer Engineering, Seoul National University, Seoul, Korea
[2]Idiap Research Institute, Martigny, Switzerland
[3]Adobe Research, San Jose, CA, USA
{mysmilesh, wanted1007, kjung}@snu.ac.kr, subhadeep.dey@idiap.ch

ABSTRACT

In this work, we explore the impact of visual modality in addition to speech and text for improving the accuracy of the emotion detection system. The traditional approaches tackle this task by independently fusing the knowledge from the various modalities for performing emotion classification. In contrast to these approaches, we tackle the problem by introducing an attention mechanism to combine the information. In this regard, we first apply a neural network to obtain hidden representations of the modalities. Then, the attention mechanism is defined to select and aggregate important parts of the video data by conditioning on the audio and text data. Furthermore, the attention mechanism is again applied to attend the essential parts of speech and textual data by considering other modalities. Experiments are performed on the standard IEMOCAP dataset using all three modalities (audio, text, and video). The achieved results show a significant improvement of 3.65% in terms of weighted accuracy compared to the baseline system.

Index Terms— speech emotion recognition, computational paralinguistics, deep learning, natural language processing

1. INTRODUCTION

Emotion identification plays a vital role in all human communications. Our response depends on emotional states, and it also provides nonverbal cues while conveying the message. Recently, there have been many efforts to understand emotions in computer-assisted technologies [1]. Automated customer service systems are typically designed to classify the emotions of the speaker to enhance the user experience. Emotion classification is also beneficial in the paralinguistic field. Recently, commercial digital assistant applications, such as Siri, found that paralinguistic information, such as emotion, is beneficial for recognizing the intent of the speaker [2]. Humans usually employ multimodality information to identify emotions. However, the impact of multimodality for this area has not yet been fully investigated. In this paper, we are interested in employing information from textual, speech, and visual models for recognizing emotions.

In the past, different approaches for emotion identification have been explored from speech signals [3, 4]. Most speech emotion techniques have focused on extracting low- or high-level features. In particular, signal processing techniques are applied to extract features, such as cepstral and prosodic features. Suprasegmental features (such as cepstral or prosodic contours) have been shown to provide good performance for this task [5]. Furthermore, statistical modeling techniques, such as the hidden Markov model (HMM) and the Gaussian mixture model (GMM), have been successfully used for this task [6, 7].

Recently, researchers have explored the application of textual information in addition to speech signals for emotion classification. Lexical information is typically used to search for keywords that express the emotional state of the speaker. In [8], lexical information is used by using a bag-of-words representation of the message. Recent approaches exploit the powerful modeling capabilities of the deep neural network (DNN) for fusing information from both modalities [9, 10, 11]. The hidden representations of the modalities are used to combine the knowledge from acoustic and textual data for emotion classification. In our previous work [12], we explored an attention mechanism for exploiting textual information. The attention mechanism is trained to summarize the lexical content and speech utterance automatically. Experimental evaluation indicates superior performance on a standard IEMOCAP dataset [13]. In this paper, we extend this approach by incorporating visual information into the framework and proposing an attention mechanism to exploit multimodal knowledge effectively. This is motivated by the fact that humans express emotion through facial expressions, speech signals, and lexical content. We hypothesize that exploiting visual knowledge in addition to speech and lexical information will result in superior performance.

As opposed to combining information from modalities independently, we propose to apply an attention mechanism that aggregates knowledge from one modality conditioned on the other two modalities. The proposed approach first obtains the sequence of hidden representations from the three modalities (speech, text, and visual). Then, the summary vector of the visual model is obtained by linearly combining the attention weights with the hidden units. Furthermore, this vector is then applied to obtain the attention weights of the acoustic data. The updated visual and acoustic data are consecutively used to compute the attention weight of the textual data to aggregate salient parts of the text modality. As this process continues multiple times, we hypothesize that the mechanism will effectively compute the relevant parts of each modality.

To evaluate the performance of the proposed approach, we performed emotion recognition experiments on the standard IEMO-CAP dataset. Experiments on this corpus indicate that the proposed approach outperforms the baseline system by 3.65% relative improvement in terms of weighted accuracy. In addition, we obtain improved model performance by increasing the iterations over the modality. The experimental results demonstrate that our proposed model correctly learns to aggregate the necessary information among the modalities via the iterative hopping process.

2. RECENT WORKS

Recently, several neural network approaches have been successfully applied for emotion classification. Researchers have proposed convolutional neural network (CNN)-based models that are trained on speech utterances for performing identification [14, 15]. There have been some successful approaches using attention mechanisms as well [16, 17]. In particular, the work in [17] presents an approach to incorporate attention-based modeling in the recurrent neural network (RNN) architecture. The attention mechanism is designed to compute the weights or relevance of each frame. An utterance level representation is obtained by temporal aggregation of these weighted speech features. The attention unit is designed to derive the segments of speech for emotion recognition automatically.

Emotion recognition using acoustic and lexical knowledge has also been explored in the literature. These works have been inspired by the fact that emotional dialogue is composed of not only speech but also textual content. In [18], emotional keywords are exploited to effectively identify the classes. Recently in [9, 10, 19], a long short-term memory (LSTM) based network has been explored to encode the information of both modalities. Furthermore, there have been some attempts to fuse the modalities using the inter-attention mechanism [11, 12]. However, these approaches are designed only to consider the interaction between acoustic and textual information.

3. MODEL

This section describes the methodologies that are applied to the speech emotion recognition task. We start by introducing a recurrent encoder to individually encode the audio, text, and video modalities. We then propose an approach to exploit each modality one by one. In this technique, an attentive modality hopping process is proposed to obtain relevant parts of each modality via the iterative aggregation process.

3.1. Recurrent Encoder

Motivated by the architecture used in [9, 17, 20], we employ a recurrent neural network to encode a series of features in the speech signal and to classify the signal into one of the emotion classes. In particular, we employ a gated recurrent unit (GRU) [21] for each modality (i.e., acoustic, textual, visual) to encode the information, as shown in figure 1. The GRU encodes the sequence of the feature vector of each modality by updating its hidden states as follows:

$$\mathbf{h}_t = f_\theta(\mathbf{h}_{t-1}, \mathbf{x}_t), \qquad (1)$$

where f_θ is the GRU network with weight parameter θ, $\mathbf{h}_t$ represents the hidden state at the t-th time step, and $\mathbf{x}_t$ represents the t-th sequential features in a target modality. This recurrent encoder is used in the same manner as the independent audio, text, and video modalities. For the video data, we obtain a fixed dimensional representation of each frame from a pretrained ResNet [22].

3.2. Proposed Attentive Modality Hopping Mechanism

We propose a novel iterative attention process, referred to as an attentive modality hopping mechanism (**AMH**), that aggregates the salient information over each modality to predict the emotion of the speech. Figure 1 shows the architecture of the proposed **AMH** model. Previous research used multimodal information independently using a neural network model by fusing information over each modality [9, 19]. Recently, researchers also investigated an

inter-attention mechanism over the modality [11, 12]. As opposed to this approach, we propose a neural network architecture that aggregates information in a modality by conditioning on other modalities via an iterative process.

First, the sequential features of each modality are encoded using the recurrent encoder by equation (1). Then, the last-step hidden states of the audio recurrent encoder, $\mathbf{h}_{\text{last}}^A$, and the text recurrent encoder, $\mathbf{h}_{\text{last}}^T$, are fused together to form contextual knowledge, $\mathbf{C}$. We then apply an attention mechanism to the video sequence, $\mathbf{h}_t^V$, to aggregate the salient part of the video modality. As this model is developed with a single attention method, we refer to the model as **AMH-1**. The final result of the **AMH-1** model, $\mathbf{H}_{\text{hop1}}$, is calculated as follows:

$$
\begin{aligned}
\mathbf{H}_{\text{hop1}} &= [\; \mathbf{h}_{\text{last}}^A; \mathbf{h}_{\text{last}}^T; \mathbf{H}_1^V \;], \\
\mathbf{H}_1^V &= \sum_i a_i \; \mathbf{h}_i^V, \\
a_i &= \frac{\exp(\; (\mathbf{C})^\mathsf{T} \mathbf{W} \mathbf{h}_i^V \;)}{\sum_i \exp(\; (\mathbf{C})^\mathsf{T} \mathbf{W} \mathbf{h}_i^V \;)}, \; (i = 1, ..., t), \\
\mathbf{C} &= f(\; \mathbf{h}_{\text{last}}^A, \mathbf{h}_{\text{last}}^T \;),
\end{aligned}
\qquad (2)
$$

where f is a fusion function (we use a vector-concatenation in this study), and $\mathbf{W} \in \mathbb{R}^{d \times d}$ and bias $\mathbf{b}$ are learned model parameters. The overall flow is presented in figure 1(a).

The $\mathbf{H}_1^V$ in equation (2) is a new modality representation for visual information considering the audio and text modality. With this information, we apply the 2nd attentive modality hopping process, referred to as **AMH-2**, to the audio sequence. The final result of the **AMH-2** model, $\mathbf{H}_{\text{hop2}}$, is calculated as follows:

$$
\begin{aligned}
\mathbf{H}_{\text{hop2}} &= [\; \mathbf{H}_1^A; \mathbf{h}_{\text{last}}^T; \mathbf{H}_1^V \;], \\
\mathbf{H}_1^A &= \sum_i a_i \; \mathbf{h}_i^A, \\
a_i &= \frac{\exp(\; (\mathbf{C})^\mathsf{T} \mathbf{W} \mathbf{h}_i^A \;)}{\sum_i \exp(\; (\mathbf{C})^\mathsf{T} \mathbf{W} \mathbf{h}_i^A \;)}, \; (i = 1, ..., t), \\
\mathbf{C} &= f(\; \mathbf{h}_{\text{last}}^T, \mathbf{H}_1^V \;),
\end{aligned}
\qquad (3)
$$

where $\mathbf{H}_1^A$ is a new representation for audio information considering the textual and visual modality after the **AMH-1** process.

We further apply the 3rd attentive modality hopping process to the textual sequence, referred to as **AMH-3**, with the updated audio and visual representations from the equations (2) and (3). The final result of the **AMH-3** model, $\mathbf{H}_{\text{hop3}}$, is calculated as follows:

$$
\begin{aligned}
\mathbf{H}_{\text{hop3}} &= [\; \mathbf{H}_1^A; \mathbf{H}_1^T; \mathbf{H}_1^V \;], \\
\mathbf{H}_1^T &= \sum_i a_i \; \mathbf{h}_i^T, \\
a_i &= \frac{\exp(\; (\mathbf{C})^\mathsf{T} \mathbf{W} \mathbf{h}_i^T \;)}{\sum_i \exp(\; (\mathbf{C})^\mathsf{T} \mathbf{W} \mathbf{h}_i^T \;)}, \; (i = 1, ..., t), \\
\mathbf{C} &= f(\; \mathbf{H}_1^A, \mathbf{H}_1^V \;),
\end{aligned}
\qquad (4)
$$

where $\mathbf{H}_1^T$ is an updated representation of the textual information considering the audio and visual modalities. Similarly, we can repeat the **AMH-1** process with updated modalities, $\mathbf{H}_1^A$, $\mathbf{H}_1^T$ and $\mathbf{H}_1^V$, to
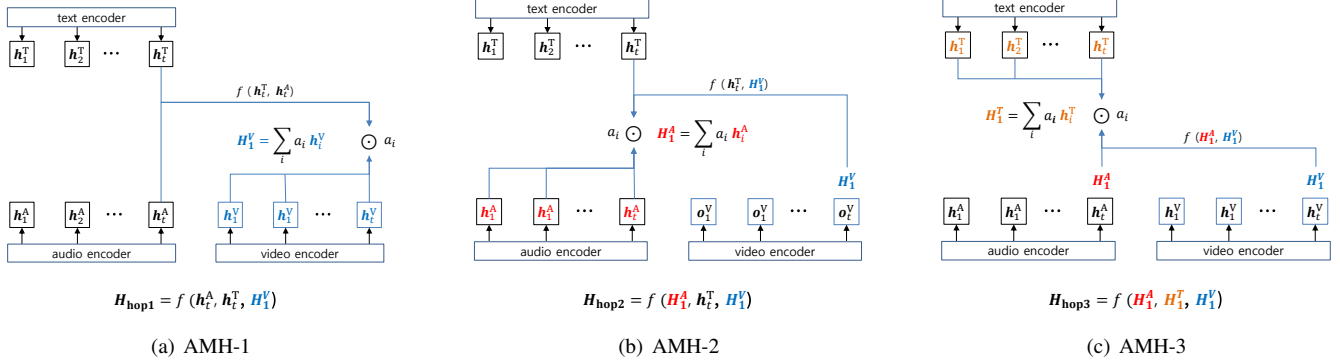
$$\mathbf{H_{hop1}} = f\left(\mathbf{h}_t^A, \mathbf{h}_t^T, \mathbf{H}_1^V\right)$$

(a) AMH-1

$$\mathbf{H_{hop2}} = f\left(\mathbf{H}_1^A, \mathbf{h}_t^T, \mathbf{H}_1^V\right)$$

(b) AMH-2

$$\mathbf{H_{hop3}} = f\left(\mathbf{H}_1^A, \mathbf{H}_1^T, \mathbf{H}_1^V\right)$$

(c) AMH-3

Fig. 1. Architecture of attentive modality-hop mechanism

define the **AMH-4** process and compute $\mathbf{H}_{hop4}$ as follows:

$$\mathbf{H}_{hop4} = [\mathbf{H}_1^A; \mathbf{H}_1^T; \mathbf{H}_2^V],$$
$$\mathbf{H}_2^V = \sum_i a_i \mathbf{h}_i^V,$$
$$a_i = \frac{\exp\left((\mathbf{C})^\mathsf{T} \mathbf{W} \mathbf{h}_i^V\right)}{\sum_i \exp\left((\mathbf{C})^\mathsf{T} \mathbf{W} \mathbf{h}_i^V\right)}, \ (i = 1, ..., t),$$
$$\mathbf{C} = f(\mathbf{H}_1^A, \mathbf{H}_1^T).$$

(5)

As the proposed **AMH** mechanism employs an iterative modality hopping process, we can generalize the **AMH-N** formulation that allows the model to hop N-times over the modality by repeating the equations (2), (3), and (4) in order.

3.3. Optimization

Because our objective is to classify speech emotion, we pass the final result of **AMH-N**, $\mathbf{H}_{hopN}$, through the softmax function to predict the seven-category emotion class. We employ the cross-entropy loss function as defined by:

$$\hat{y}_c = \mathrm{softmax}((\mathbf{H}_{hopN})^\mathsf{T} \mathbf{W} + \mathbf{b}),$$
$$\mathcal{L} = \frac{1}{N} \sum_{j=1}^N \sum_{c=1}^C y_{j,c} \log(\hat{y}_{j,c}),$$

(6)

where $y_{j,c}$ is the true label vector, and $\hat{y}_{j,c}$ is the predicted probability from the softmax layer. The $\mathbf{W}$ and the bias $\mathbf{b}$ are model parameters. C is the total number of classes, and N is the total number of samples used in training.

4. EXPERIMENTS

4.1. Dataset and Experimental Setup

We use the Interactive Emotional Dyadic Motion Capture (IEMO-CAP) [13] dataset, which contains abundant multimodal emotion descriptions of natural utterances. The corpus includes five sessions of utterances between two speakers (one male and one female). A total of 10 unique speakers participated in this work. The emotional category for each utterance was annotated by three people. First, we eliminated all data that were labeled as three different emotions. Then, following previous research [9, 10, 12], we assign a single categorical emotion to the utterance in which the majority of annotators agreed on the emotion labels. The final dataset contains 7,487

utterances in total (1,103 *angry*, 1,041 *excite*, 595 *happy*, 1,084 *sad*, 1,849 *frustrated*, 107 *surprise* and 1,708 *neutral*). We do not include classes from the original dataset that is too small in size, such as 3 *other*, 40 *fear* and 2 disgust. In the training process, we perform 10-fold cross-validation where each 8-, 1-, and 1-fold are used for the training set, development set, and test set, respectively.

4.2. Feature extraction and Implementation details

As this research is extended work from previous research [12], we use the same feature extraction method for audio and text modality as in our previous work. After extracting the 40-dimensional mel-frequency cepstral coefficients (MFCC) feature (frame size is set to 25 ms at a rate of 10 ms with the Hamming window) using Kaldi [23], we concatenate it with its first- and second-order derivatives, making the feature dimension 120.

In preparing the textual dataset, we first use the ground-truth transcripts of the IEMOCAP dataset. We further obtain the transcripts by using a commercial automatic speech recognition system [24] (the performance of the ASR system is 5.53% word error rate) for the practical use case where the ground-truth may not be available. For both cases, we apply a word-tokenizer to the transcription and obtain sequential features of the textual modality.

For the visual dataset, we perform an additional prepossessing. The video data in the IEMOCAP are recorded with the two actors together in the video frame. We first split each video frame into two subframes so that each segment contains only one actor. Then, we crop the center of each frame with size 224*224 to focus on the actor and to remove the background in the video frame. Finally, we extract 2,048-dimensional visual features from each subframe using pretrained ResNet-101 [22] at a frame rate of 3 per second.

We minimize the cross-entropy loss function (equation (6)) using the Adam optimizer [25] with a learning rate of 1e-3 and gradients clipped with a norm value of 1. All of the code developed for the empirical results is available via web repository.[1]

4.3. Performance evaluation

We report the model performance using the weighted accuracy (WA) and unweighted accuracy (UA). We perform 10-fold cross-validation (10 experiments for each fold) and report the average and standard deviation results.

Table 1 shows the model performance on the speech emotion recognition task. To compare our results with previous approaches,

---

[1]http://github.com/david-yoon/attentive-modality-hopping-for-SER

Model	Modality	WA	UA
	Ground-truth transcript		
RNN-attn [17]	A	0.359 ±0.058	0.319 ±0.065
RNN-attn [17]	T	**0.528** $\pm$**0.018**	**0.525** $\pm$**0.031**
RNN-attn [17]	V	0.467 ±0.046	0.412 ±0.066
MDRE [9]	A+T	0.557 ±0.018	0.536 ±0.030
MDRE [9]	T+V	0.585 ±0.040	0.561 ±0.046
MDRE [9]	A+V	0.481 ±0.049	0.415 ±0.047
MHA [12]	A+T	0.583 ±0.025	0.555 ±0.040
MHA [12]	T+V	**0.590** $\pm$**0.017**	**0.560** $\pm$**0.032**
MHA [12]	A+V	0.490 ±0.049	0.434 ±0.060
MDRE [9]	A+T+V	0.602 ±0.033	0.575 ±0.046
AMH (ours)	A+T+V	**0.624** $\pm$**0.022**	**0.597** $\pm$**0.040**
	ASR-processed transcript		
AMH-*ASR* (ours)	A+T+V	0.611 ±0.024	0.595 ±0.036

Table 1. Model performance comparisons. The "A", "T" and "V" in modality indicate "Audio", "Text" and "Video", receptively. The "-*ASR*" models use ASR-processed transcripts.

# hop	context		target	WA	UA
1	$\mathbf{h}_{last}^A$	$\mathbf{h}_{last}^T$	V	0.599 ±0.015	0.579 ±0.032
2	$\mathbf{h}_{last}^T$	$\mathbf{H}_1^V$	A	0.610 ±0.024	0.589 ±0.047
3	$\mathbf{H}_1^A$	$\mathbf{H}_1^V$	T	0.605 ±0.033	0.581 ±0.051
4	$\mathbf{H}_1^A$	$\mathbf{H}_2^T$	V	**0.612** $\pm$**0.030**	**0.586** $\pm$**0.041**
5	$\mathbf{H}_2^T$	$\mathbf{H}_2^V$	A	0.600 ±0.038	0.581 ±0.054
6	$\mathbf{H}_2^A$	$\mathbf{H}_2^V$	T	0.610 ±0.023	0.584 ±0.046
7	$\mathbf{H}_2^A$	$\mathbf{H}_3^T$	V	**0.624** $\pm$**0.022**	**0.597** $\pm$**0.040**
8	$\mathbf{H}_3^T$	$\mathbf{H}_3^V$	A	0.603 ±0.034	0.576 ±0.047
9	$\mathbf{H}_3^A$	$\mathbf{H}_3^V$	T	0.608 ±0.034	0.583 ±0.048

Table 2. Model (**AMH**) performance as the number of hop increases. Top-2 scores marked in bold.

Fig. 2. Confusion matrix of the AMH-7 model

we report the model performance in regards to the types of modalities used for the experiments. From the previous model, **RNN-attn** employs attention memory to compute and aggregate the emotionally salient part while encoding a single modality using a bidirectional LSTM network. It achieves the best performance of 0.528 WA, with the use of the textual modality. In contrast, the **MDRE** model use multiple RNNs to encode multiple modalities and merge the results using another fully connect neural network layer. Similarly, the **MHA** model employs dual-RNN for any two modalities and compute inter-attention over each modality. Among the models that use any two modalities, **MHA** achieves the best performance of 0.590 WA, with the textual and visual modalities.

Finally, we evaluate model performances with the use of three modalities, audio, text, and video. Our proposed model, **AMH**, outperforms **MDRE** by 3.65% relative (0.602 to 0.624 absolute) in terms of WA. Note that we omit the **MHA** experiments with three modalities since they cannot deal with three modalities. In a practical scenario, we may not access the audio transcripts. We describe the effect of using ASR-processed transcripts on the proposed system. As shown in table 1, we observe performance degradation in **AMH-ASR** compared to that of **AMH** (our best system) by 2.08% (0.624 to 0.611) relative in WA. Even with the erroneous transcripts, however, the proposed system **AMH-ASR** surpasses the baseline system (**MDRE**) by 1.49% relative (0.602 to 0.611) in terms of WA.
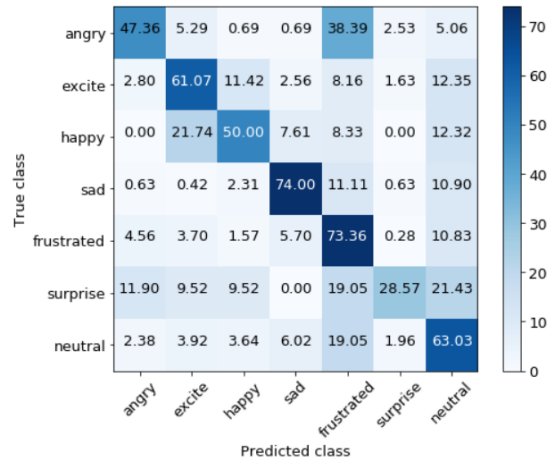
Table 2 shows the model performance (**AMH**) as the number of hops increases. As the number of hops increases, the representative vector of each modality gets updated in order according to equation (2)-(4). We find that the model achieves the best performance in the 7-hop case. This behavior demonstrates that our proposed model correctly learns to aggregate the salient information among the modalities via the iterative hopping process.

4.4. Error analysis

Figure 2 shows the confusion matrices of the proposed systems. In general, most of the emotion labels are frequently misclassified as *neutral* class, supporting the claims of [12, 26]. The model confused the *excite* and *happy* classes since there exists a report of overlap in distinguishing these two classes, even in human evaluations [13]. It is interesting to observe that the model misclassifies the *angry* class to the *frustrated* class with a rate of 0.383; however, the rate is 0.045 in the opposite case (confusion *frustrated* to *angry* class). It is natural to see that the model is the most inaccurate in the *surprise* class because we only have a small-size dataset for that class (107 samples).

5. CONCLUSIONS

In this paper, we proposed an attentive modality hopping mechanism for speech emotion recognition tasks. The proposed mechanism computes the salient part and aggregates the sequential information from one modality by conditioning on the other two modalities via an iterative hopping process. Extensive experiments demonstrate that the proposed **AMH** surpasses the best baseline system by 3.65% relative improvement in terms of weighted accuracy.

Acknowledgments

6. REFERENCES

[1] Rosalind W Picard, "Affective computing: challenges," *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 55–64, 2003.

[2] Vikramjit Mitra, Sue Booker, Erik Marchi, David Scott Farrar, Ute Dorothea Peitz, Bridget Cheng, Ermine Teves, Anuj Mehta, and Devang Naik, "Leveraging acoustic cues and paralinguistic embeddings to detect expression from voice," *arXiv preprint arXiv:1907.00112*, 2019.

[3] Kun Han, Dong Yu, and Ivan Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Fifteenth annual conference of the international speech communication association*, 2014.

[4] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.

[5] Weihui Dai, Dongmei Han, Yonghui Dai, and Dongrong Xu, "Emotion recognition and affective computing on vocal social media," *Information & Management*, vol. 52, no. 7, pp. 777–788, 2015.

[6] Björn Schuller, Gerhard Rigoll, and Manfred Lang, "Hidden markov model-based speech emotion recognition," in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03)*. IEEE, 2003, vol. 2, pp. II–1.

[7] Moataz MH El Ayadi, Mohamed S Kamel, and Fakhri Karray, "Speech emotion recognition using gaussian mixture vector autoregressive models," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. IEEE, 2007, vol. 4, pp. IV–957.

[8] Kalani Wataraka Gamage, Vidhyasaharan Sethu, and Eliathamby Ambikairajah, "Salience based lexical features for emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5830–5834.

[9] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 112–118.

[10] Jaejin Cho, Raghavendra Pappagari, Purva Kulkarni, Jesús Villalba, Yishay Carmiel, and Najim Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," *Proc. Interspeech 2018*, pp. 247–251, 2018.

[11] Haiyang Xu, Hui Zhang, Kun Han, Yun Wang, Yiping Peng, and Xiangang Li, "Learning alignment for multimodal emotion recognition from speech," *Proc. Interspeech 2019*, pp. 3569–3573, 2019.

[12] Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung, "Speech emotion recognition using multi-hop attention mechanism," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2822–2826.

[13] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[14] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Platform Technology and Service (PlatCon), 2017 International Conference on*. IEEE, 2017, pp. 1–5.

[15] Zakaria Aldeneh and Emily Mower Provost, "Using regional saliency for speech emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2741–2745.

[16] Pengcheng Li, Yan Song, Ian McLoughlin, Wu Guo, and Lirong Dai, "An attention pooling based representation learning method for speech emotion recognition," *Proc. Interspeech 2018*, pp. 3087–3091, 2018.

[17] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2227–2231.

[18] Björn Schuller, Gerhard Rigoll, and Manfred Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Acoustics, Speech, and Signal Processing (ICASSP), IEEE International Conference on*. IEEE, 2004, pp. I–577.

[19] Jilt Sebastian and Piero Pierucci, "Fusion techniques for utterance-level emotion recognition combining speech and transcripts," *Proc. Interspeech 2019*, pp. 51–55, 2019.

[20] Yun Wang, Leonardo Neves, and Florian Metze, "Audio-based multimedia event detection using deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2742–2746.

[21] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder–decoder for statistical machine translation," in *Proceedings of the EMNLP*, 2014, pp. 1724–1734.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[23] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," Tech. Rep., IEEE Signal Processing Society, 2011.

[24] Google, "Cloud speech-to-text," 2019, Accessed: 2020-02-01.

[25] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[26] Michael Neumann and Ngoc Thang Vu, "Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech," *Proc. Interspeech 2017*, pp. 1263–1267, 2017.