Multimodal and Multi-view Models for Emotion Recognition

Gustavo Aguilar[‡], **Viktor Rozgić**[†], **Weiran Wang**[†] and **Chao Wang**[†]
University of Houston[‡]
Amazon.com[†]
gaguilaralas@uh.edu, {rozgicv, weiranw, wngcha}@amazon.com

Abstract

Studies on emotion recognition (ER) show that combining lexical and acoustic information results in more robust and accurate models. The majority of the studies focus on settings where both modalities are available in training and evaluation. However, in practice, this is not always the case; getting ASR output may represent a bottleneck in a deployment pipeline due to computational complexity or privacy-related constraints. To address this challenge, we study the problem of efficiently combining acoustic and lexical modalities during training while still providing a deployable acoustic model that does not require lexical inputs. We first experiment with multimodal models and two attention mechanisms to assess the extent of the benefits that lexical information can provide. Then, we frame the task as a multi-view learning problem to induce semantic information from a multimodal model into our acoustic-only network using a contrastive loss function. Our multimodal model outperforms the previous state of the art on the USC-IEMOCAP dataset reported on lexical and acoustic information. Additionally, our multi-view-trained acoustic network significantly surpasses models that have been exclusively trained with acoustic features.

1 Introduction

The task of emotion recognition (ER) requires understanding the way humans interact to express their emotional state during conversations. Among others, emotions are encoded in both lexical and acoustic information where each modality contributes to the overall emotional state of a given speaker. However, in some situations, one modality can be more insightful to derive emotions than the other. For instance, the phrase *"yeah... of course"* does not have enough lexical information to derive the right emotion, and it may all depend on the acoustic patterns. On the other hand, the phrase *"I really miss my dog!"* does not need acoustic information to detect that the most likely emotion is sadness. Thus, recognizing emotions is not a trivial task because an emotional state can be easily shaped by many factors: context, word content, spectral and prosodic information, among others (Barbulescu et al., 2017).

In this paper, we study the emotion recognition problem from the speech and language perspectives. We formally look into acoustic and lexical modalities with the aim of improving models that only use acoustic information. In the first part of this work, our goal is to assess the extent to which lexical information benefits acoustic models. We propose a multimodal method that is inspired by the way humans process emotions in a conversation. That is, lexical and acoustic information is simultaneously perceived at every word step. Hence, we introduce the concept of acoustic words: word-level representations derived from acoustic features in a speech fragment. The acoustic word representations enable a natural combination of the modalities where lexical and acoustic features are aligned at the word level. Additionally, we leverage these representations with two attention mechanisms: modality-based and context-based attentions. The former mechanism prioritizes one of the modalities at each word step, whereas the latter mechanism focuses on the most important word representations across the entire utterance. Our multimodal approach outperforms the current state of the art on the USC-IEMOCAP dataset reported on lexical and acoustic modalities.

In the second part of this work, our goal is to induce semantic information from the proposed multimodal model into an acoustic model. We study a more challenging scenario where we establish that lexical information is available during

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 991–1002
Florence, Italy, July 28 - August 2, 2019.

training but not during the evaluation phase. Such restriction is commonly found in real-world applications, where transcripts or ASR outputs represent a bottleneck in a deployment pipeline due to computational complexity or privacy-related constraints. To address this challenge, we frame this task as a multi-view learning problem (Blum and Mitchell, 1998). We induce lexical information from our multimodal model into the acoustic network during training while still providing a lexical-independent acoustic model for testing or deployment. That is, our acoustic model learns to capture semantic and contextual information without relying on explicit lexical inputs such as ASR or transcripts. This multi-view acoustic network significantly outperforms models that have been exclusively trained on acoustic features.

2 Related Work

Recognizing emotions is a complex task because it involves several ambiguous human interactions such as facial expressions, change in pitch or tone of voice, linguistic semantics and meaning, among others (Cowie, 2009; Mower Provost et al., 2009). Many researchers have approached these challenges by extracting features from visual, acoustic, and lexical information. Early approaches rely on a variation of support vector machine (SVM) classifiers to learn emotional categories such as happiness, sadness, anger, and others (Rozgic et al., 2012; Perez-Rosas et al., 2013; Jin et al., 2015). For instance, Rozgic et al. (2012) use an automatically generated ensemble of trees whose nodes contain binary SVM classifiers for each emotional category. Jin et al. (2015) also use multimodality, and their study focuses on comparing early and late-fusion methods. Consistently, researchers have found that multimodal approaches outperform unimodal ones.

Recent work has focused on different ways to fuse the acoustic, lexical, and visual modalities. However, we narrow the discussion to the acoustic and lexical modalities to align with the scope of the paper. In most of the cases, researchers have used concatenation to fuse the lexical and acoustic representations at different stages of their models. Other works have proposed multimodal pooling fusion (Aldeneh et al., 2017), tensor fusion networks (Zadeh et al., 2017), modality hierarchical fusion (Majumder et al., 2018), context-aware fusion with attention (Poria et al., 2018), and conversational memory networks (CMN) (Hazarika et al., 2018). Nevertheless, all the previous fusion techniques have been made at the utterance level, whereas our work focuses on multimodal fusion at the word level by introducing acoustic word representations. We compare our work to Poria et al. (2018) because they document the current best performance on lexical and acoustic information on the IEMOCAP dataset using the standard 10-fold speaker-exclusive cross-validation setting.

Closely related work on acoustic word embeddings has been made by He et al. (2016). They induce acoustic information into lexical representations at the character level in a multi-view unsupervised setting. We introduce the concept of acoustic word representations in a different way: we learn vector representations of words out of frame-level acoustic features. This allows us to align lexical and acoustic information at the word level, which simulates the way humans perceive emotions in conversations (i.e., both modalities are simultaneously perceived).

We also explore multi-view settings to overcome the absence of lexical inputs during evaluation (Blum and Mitchell, 1998). There are multiple options to conduct the experiments in this scenario (Xu et al., 2013; Wang et al., 2015), such as deep cannonical correlation analysis (DCCA) (Andrew et al., 2013) and siamese networks with contrastive loss functions (He et al., 2016). We use the latter approach in our experiments. To the best of our knowledge, there is no prior work trying to overcome the absence of lexical inputs by inducing lexical information into an acoustic model for the task of emotion recognition.

3 Methodology

We describe the data representation and introduce the idea of acoustic words in Sections 3.1 and 3.2. Then, we use this concept to define the multimodal architecture in Section 3.3. Finally, we explain the multi-view learning setting using the proposed multimodal model and our acoustic model in Section 3.3.

3.1 Data Representation

Acoustic features. We extract frame-level features using OpenSMILE[1] (Eyben et al., 2013). We use the Computational Paralinguistic Challenge (ComParE) feature-set introduced by Schuller

[1] audeering.com/technology/opensmile/

et al. (2013) for the InterSpeech emotion recognition challenge. These features include energy, spectral, MFCC, and other low-level descriptors. The InterSpeech ComParE 2013 features are fairly standard and well-documented. Additionally, we normalize these features using z-standardization before feeding them into our models.

Lexical features. We use word embeddings to represent the lexical information. Specifically, we employ deep contextualized word representations using the language model ELMo (Peters et al., 2018). ELMo represents words as vectors that are entirely built out of characters. This allows us to overcome the problem of out-of-vocabulary words by always having a vector based on morphological clues for any given word. Additionally, these representations have proven to capture syntax and semantics aspects as well as the diversity of the linguistic context of words (e.g., polysemy).

3.2 Acoustic Words

Previous studies usually extract features from the modalities in independent modules, and then they concatenate the corresponding utterance representations from the acoustics and lexical features to feed into the next layers of their models. However, we argue that a more natural way to understand emotions is to align lexical and acoustic information, which simulates the way humans process both modalities simultaneously. Thus, we introduce the concept of acoustic word representations (see Figure 1). These representations are extracted from frame-level features by taking the output of a bidirectional LSTM at every segment. Note that this procedure requires the word alignment information. Additionally, we exclude frames that do not belong to the words of the speaker. This reduces any potential bias towards other people's emotional states as well as environmental noise.

3.3 Hierarchical Multimodal Model

Our goal is to provide a neural network model that efficiently combines acoustic and lexical information for emotion recognition. We propose a hierarchical multimodal model that uses: 1) acoustic word representations derived from frame-level features, 2) a modality-based attention mechanism at the word level that prioritizes one modality over the other, and 3) a context-based attention mechanism that emphasizes the most relevant parts in the entire utterance. In Figure 1, the shadowed box represents the low level of the hierarchy, where the frame features are used to generate the acoustic word representation. The high level of the model is where the word representations from each modality are combined.

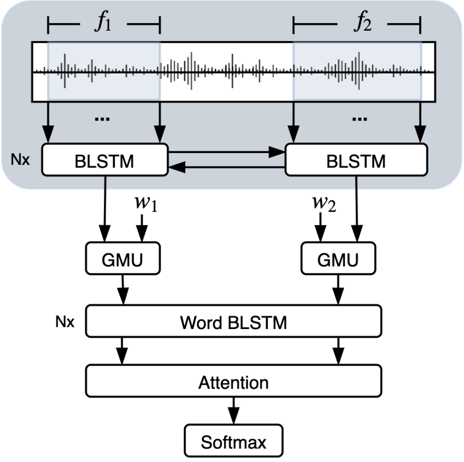

Figure 1: The multimodal model. The shadowed box incloses the acoustic word mechanism, whose output is fed into the GMU unit along with the lexical word representation at each timestep. The model can have N layers of BLSTM at the frame and word levels.

Modality-based attention. The idea of the modality-based attention is to prioritize one of the modalities at the word level. That is, when the lexical features are more relevant to capture emotions (i.e., informative words are used), the model should prioritize such features and vice versa (i.e., arousal and pitch levels increase). To achieve this behavior, we incorporate the bimodal version of the GMU cell proposed by Arevalo et al. (2017). The GMU equations are as follows:

$$
\begin{aligned}
h_a &= \tanh(\mathrm{W}_a x_a + b_a) \\
h_l &= \tanh(\mathrm{W}_l x_l + b_l) \\
z &= \sigma(\mathrm{W}_z [x_a, x_l] + b_z) \\
h &= z * h_a + (1 - z) * h_l
\end{aligned}
\tag{1}
$$

where x_a and x_l are the acoustic and lexical input vectors, respectively. These inputs are concatenated ($[x_a, x_l]$) and then multiplied by W_z so that the concatenation can be projected into the same space of the hidden vectors h_a and h_l. Finally, z is multiplied by the hidden acoustic vector h_a, and $(1 - z)$ by the hidden lexical vector h_l. By adding the result of these products, the model incorporates a complementary mechanism over the modalities, which allows prioritizing one over the other when necessary.

Context-based attention. We use a fairly standard attention mechanism over the entire utterance

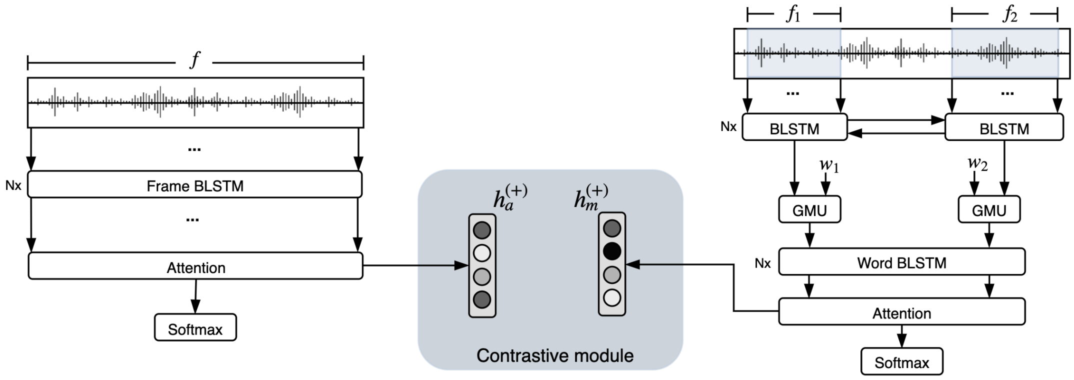

Figure 2: The multi-view models. The view on the left is the acoustic model, and the view on the right is the multimodal model. The shadowed box in the middle is the contrastive loss module.

that was introduced by Bahdanau et al. (2014). The idea is to concentrate mass probability over the words that capture emotional states along the sequence. Our attention mechanism uses the following equations:

$$e_i = v^{\intercal} \tanh(\mathrm{W}_h h_i + b_h)$$

$$a_i = \frac{\exp(e_i)}{\sum_{j=1}^{N} \exp(e_j)}, \quad \text{where} \sum_{i=1}^{N} a_i = 1$$

$$z = \sum_{i=1}^{N} a_i h_i$$

where $\mathrm{W}_h \in \mathbb{R}^{d_a \times d_h}$ and $b_h \in \mathbb{R}^{d_h}$ are trainable parameters of the model. The vector $v \in \mathbb{R}^{d_a}$ is the attention vector to be learned. Also, d_a and d_h are the dimensions of the attention layer and the hidden state, respectively. Then, we multiply the scalars a_i and their corresponding hidden vectors h_i to obtain our weighted sequence. The sum of the weighted vectors, z, is fed into a softmax layer.

3.4 Multi-view Learning

A more realistic and challenging scenario happens when lexical information is not available during testing. In this case, our goal is to build an acoustic model that is capable of inferring some notion of semantic and contextual features by taking advantage of lexical information only available during training. To achieve this, we frame the problem as a multi-view learning task, where two disjoint networks share their learned information through the loss function (Lian et al., 2018). The fact that they are disjoint networks allows them to function without each other during evaluation.

Consider the acoustic and multimodal views V_a and V_m. The acoustic view, V_a, is comprised of N layers of bidirectional LSTMs followed by an attention and a softmax layers. The multimodal view, V_m, follows the architecture described in Section 3.3. As shown in Figure 2, the view on the left, V_a, takes only the raw frame vectors, whereas the view on the right, V_m, takes the aligned frame and word vectors as inputs. Each view learns an utterance representation of the emotions, h_a and h_m, which are the outputs of their corresponding attention layers, as defined in Eq. 2. Since these vectors come from the same source of information (i.e., same speaker utterance), we assume that their emotion representations are similar. In general, we want vectors with similar emotions to be close and dissimilar ones to be far regardless of the modalities they use. To achieve this, we use the following contrastive loss function:

$$\begin{aligned}\mathcal{L}_c =& \frac{1}{2N}\sum_{i}^{N} \max(0, m + dis(h_{a_i}, h_{m_i}^{+}) - dis(h_{a_i}, h_{m_i}^{-})) \\ &+ \frac{1}{2N}\sum_{i}^{N} \max(0, m + dis(h_{m_i}, h_{a_i}^{+}) - dis(h_{m_i}, h_{a_i}^{-}))\end{aligned} \quad (2)$$

where the $+$ and $-$ superscripts refer to positive (i.e., close) and negative (i.e., far) vectors. We force a margin of at least m to keep negative samples separated from positive samples. We define $dis(v, w) = 1 - cos(v, w)$ as the function that calculates the distance between two vectors. Note that we determine cross-view pairs when comparing vectors because we want the models to induce similar information from different modalities. Additionally, choosing the negative samples can dramatically affect the performance of the models. For instance, for random samples that may not share acoustic or lexical properties, the models can easily satisfy the margin m without forcing much learning. Instead, we want the models to find the nuances in acoustically similar samples that have different emotion labels. Thus, besides random sampling, we also consider similar acoustic properties (e.g., valence, arousal, or dominance) that overlap among the emotions.

In addition to the contrastive objective function, we use cross-entropy loss functions for the acoustic and multimodal views:

$$\mathcal{L}_a = -\frac{1}{N}\beta_a \sum_{i}^{N} y_i log(\hat{y}_i) \tag{3}$$

$$\mathcal{L}_m = -\frac{1}{N}\beta_m \sum_{i}^{N} y_i log(\hat{y}_i) \tag{4}$$

where β_a and β_m are used to weight the loss from the acoustic and multimodal views, respectively. These weights can vary along the epochs to facilitate the optimization of the acoustic view. We discuss this in Section 4.4, and the training procedure is described in Algorithm 1.

Algorithm 1 Multi-view Training Algorithm

```
procedure GetNegSamples($Data$, y)
    ▷ Loop through the targets of the batch
    for $i \leftarrow 1, \dots, ||y||$ do
        ▷ Randomly pick sample with class other than $y_i$
        $y_i^- \leftarrow$ Rand($Data$)
                s.t. $y_i^- \neq y_i$ and
                     $y_i^-, y_i$ are acoustically similar
        ▷ Collect the corresponding negative inputs
        $(x_{a_i}^-, x_{l_i}^-) \leftarrow getinput(y_i^-)$
    return $(x_a^-, x_l^-)$
repeat:
    ▷ Loop through the training batches
    for $(x_a, x_l, y) \leftarrow nextbatch(Data)$ do
        ▷ Get the negative acoustic and lexical inputs
        $(x_a^-, x_l^-) \leftarrow$ GetNegSamples($Data$, y)
        ▷ Get the neg. hidden vectors from neg. inputs
        $h_a^- \leftarrow hidden(V_a, x_a^-)$
        $h_m^- \leftarrow hidden(V_m, x_a^-, x_l^-)$
        ▷ Get the pos. hidden vectors and predictions
        $(h_a, \hat{y}_a) \leftarrow forward(V_a, x_a)$
        $(h_m, \hat{y}_m) \leftarrow forward(V_m, x_a, x_l)$
        ▷ Calculate and add the individual losses
        $\mathcal{L}_c \leftarrow$ Contrastive($h_a, h_a^-, h_m, h_m^-$)
        $\mathcal{L}_a \leftarrow$ CrossEntropy($y, \hat{y}_a$)
        $\mathcal{L}_m \leftarrow$ CrossEntropy($y, \hat{y}_m$)
        $\mathcal{L} \leftarrow \mathcal{L}_c + \beta_a \mathcal{L}_a + \beta_m \mathcal{L}_m$
        ▷ Update the parameters using backprop.
        $\Theta_{V_m} \leftarrow \Theta_{V_m} - \alpha \partial \mathcal{L} / \partial \Theta_{V_m}$
        $\Theta_{V_a} \leftarrow \Theta_{V_a} - \alpha \partial \mathcal{L} / \partial \Theta_{V_a}$
until stopping criteria met
```

Utterances	Anger	Happiness	Neutral	Sadness
F1 - 528	147	132	171	78
M1 - 556	82	146	212	116
F2 - 479	67	166	134	112
M2 - 542	70	161	227	84
F3 - 522	92	128	130	172
M3 - 624	148	154	190	132
F4 - 527	205	185	75	62
M4 - 501	122	118	180	81
F5 - 590	78	159	221	132
M5 - 651	92	283	163	113
5,520	1,103	1,632	1,703	1,082

Table 1: Data distribution of the USC-IEMOCAP dataset. F and M mean female and male speakers followed by their session number.

Teacher-student learning. We anticipate two potential problems with the previously described setting: 1) the learning process may predominantly concentrate on the multimodal view because it has more learning capabilities (i.e., large number of parameters) than the acoustic view, leaving the acoustic model to be of secondary importance during training, and 2) a cross-entropy loss over one-hot vectors ignores informative overlaps among the emotion classes resulting in a very strict objective function. To address these issues, we look into a teacher-student learning approach (Li et al., 2014). Given an already-optimized multimodal model V_m (the teacher), we want our acoustic view V_a (the student) to predict probability distributions such as the ones generated by the teacher. We can calculate the difference between the probability distributions of the teacher and the student using Kullback-Leibler (KL) divergence. Then, we minimize the following loss function:

$$\mathcal{L}_{KL} = -\frac{1}{N}\sum_{i}^{N} p(y_i|x_{m_i}, V_m) log \frac{p(y_i|x_{m_i}, V_m)}{p(y_i|x_{a_i}, V_a)} \tag{5}$$

where x_{m_i} and x_{a_i} are the multimodal and acoustic inputs for sample i, respectively, and V_m and V_a represent the parameters of the views.

4 Experiments

We describe the dataset used for the experiments in Section 4.1. Then, we define the experimental models in Section 4.2, which are used in the multimodal and multi-view experiment in Sections 4.3 and 4.4.

4.1 Dataset

We focus our experiments on the USC-EIMOCAP dataset (Busso et al., 2008). This dataset provides

Type	Experiment	Modality	Dev	Test	Comment
Baseline	B-ACO-1	Acoustics	0.5858	-	Silence frames
	B-ACO-2		0.5729	-	Silence frames removed
	B-LEX	Lexical	0.6706	-	-
	B-MM-1	Multimodal	0.7195	-	Silence frames
	B-MM-2		0.7265	-	Silence frames removed
Hierarchical	H-ACO-1	Acoustics	0.5697	-	Acoustic words
	H-MM-1	Multimodal	0.7316	-	Aligned words
	H-MM-2		0.7341	-	+ GMU
	H-MM-3		0.7354	-	+ Attention
	H-MM-4		**0.7383**	**0.7169**	+ GMU + Attention
SOTA	-	Multimodal	-	**0.7079**	Poria et al. (2018)

Table 2: The results of the multimodal experiments. The name of the experiments starts either with B or H referring to baseline or hierarchical models. ACO, LEX, and MM mean acoustic, lexical and multimodal. Our results provide a new state-of-the-art UA when we use the hierarchical model with GMU and attention. Once the models are optimized on the validation set, we evaluate the best ones on the test set.

conversations between female and male speakers throughout five sessions. Each session involves a different pair of speakers, which accounts for a total of 10 speakers. The conversations are split into small utterances that map to emotion categories. The original emotion categories are merged to mitigate the unbalanced classes into four categories: *anger*, *happiness*, *neutral*, and *sadness*. Table 1 shows the distribution of the dataset. We split the dataset using the one-speaker-out experimental setting. That is, we take four sessions for training, and the remaining session is split by speakers into the validation and test sets. We report our unweighted accuracy scores running 10-fold cross-validation experiments and averaging scores across folds.

4.2 Defining Experimental Models

B-ACO: The acoustic baseline is composed of two BLSTM layers of 256 dimensions each, followed by average pooling and a softmax layer. B-ACO-1 uses the raw sequence of frames, whereas B-ACO-2 employs the frames that correspond to the speaker.

B-LEX: The lexical baseline uses word embeddings of 1,024 dimensions from ELMo. We feed these vectors into two BLSTM layers of 256 dimensions followed by average pooling and a softmax layer.

B-MM: The multimodal baseline uses BLSTMs with average pooling over time on each modality, similar to B-ACO and B-LEX. We concatenate the vectors from each modality and feed them into a softmax layer.

H-ACO: The hierarchical acoustic model uses acoustic word representations. The acoustic words are generated with two BLSTMs of 256 dimensions using the speaker frames (i.e., no silence). At the word level, we perform average pooling over time and feed the resulting vector into a softmax layer.

H-MM: The hierarchical multimodal model uses the acoustic word representations in H-ACO, and the lexical word representations in B-LEX, with 256 dimensions each. H-MM-1 uses two layers of BLSTM over the concatenated word representations followed by average pooling and a softmax layer. Based on H-MM-1, H-MM-2 incorporates the GMU unit and H-MM-3 adds the attention layer. H-MM-4 uses both GMU and the attention layer.

4.3 Multimodal Experiments

Impact of silence. We experiment with silence and the baselines B-ACO and B-MM. In Table 2, although keeping silence seems better than removing it (B-ACO-1 vs. B-ACO-2), the multimodal model shows a small improvement when silence is ignored (B-MM-1 vs. B-MM-2). By looking into the predictions, besides the silence and environmental noise in the original frames, we notice that a second speaker can influence the emotions of the speaker being evaluated. This observation, along with the model improvements, suggests that

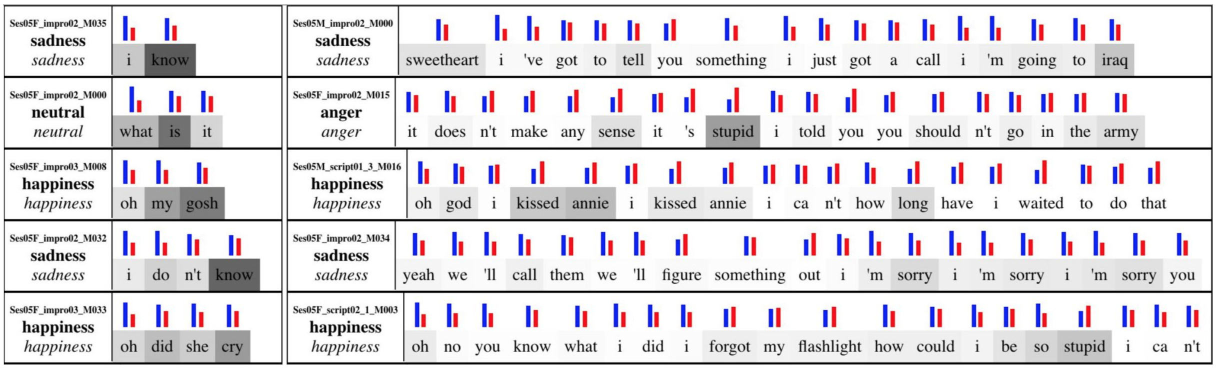

Figure 3: Multimodal Attention. The figure shows the attention mechanisms at the modality and utterance levels. The bars over the words are the average of z in Eq. 1, and they show how much acoustic (left bar in blue) or lexical (right bar in red) information was used. The highlights in the background of the words are the attention probabilities, where the higher the probability the darker the word.

is possible to fuse information more efficiently.

Hierarchical models. To make better use of the modalities, we align lexical information with acoustic representations at the word level. Based on the silence impact, our acoustic word representations only use frames where the speaker intervenes in the conversation (i.e., no silence or other speakers). Similar to the previous scenario, we see a detrimental behavior in the hierarchical acoustic model compared to the models that use the original sequence of frames (H-ACO-1 vs. B-ACO). However, when we concatenate the lexical and acoustic word representations (H-MM-1), our hierarchical model surpasses the UA of all previous models. In fact, our best model (H-MM-4) outperforms the previous state-of-the-art UA. This serves as strong evidence that fusing information more efficiently can yield a better performance.

Ablation experiment. Table 2 shows the performance of the hierarchical multimodal models with and without the modality- and context-based attention mechanisms (H-MM). Using H-MM-1 as a common ground, the modality-based attention (H-MM-2) provides an improvement of about 1% on the UA metric. This result suggests that one modality can be more informative than the other, and hence, it is important to prioritize the one that carries more emotional information. Likewise, adding the attention mechanism, H-MM-3, outperforms H-MM-1 by a similar percentage. Our intuition is that weighting the words that provide strong emotional information based on the context allows the model to disambiguate meaning and discriminate more easily the samples. Lastly, H-MM-4 combines both attention mechanisms, which improves over the individual attention models H-MM-2 and H-MM-3 by about 1% of UA. This means that the attention mechanisms are more complementary than overlapping.

Attention visualization. For the modality-based attention, the vector z from Eq. 1 determines how much acoustic information will go through the next layers, whereas $(1 - z)$ is the amount of lexical data allowed. Figure 3 provides a visualization of these vectors. The bars show the amount of information that is captured from one modality versus the other. For instance, the sample *"oh my gosh"* illustrates that the words rely on more acoustic than lexical information. Intuitively, this phrase by itself could describe different emotions, but it is the acoustic modality that mitigates the ambiguity. Regarding the context-based attention, Figure 3 shows the places where the model focuses along the utterance. For large-context utterances, where the acoustic features are more or less similar, the semantics can help to highlight specific spots. For example, in the second sentence on the right of Figure 3, the model detects the semantics of the words *sense* and *stupid* and associates them with the words *should*, *go*, and *army*. The attention mechanism not only emphasizes semantics but it also takes into account the acoustic features. In the same block of sentences, it is worth noting that the words primarily driven by acoustics (e.g., *sweatheart*, *oh god*, *sorry* and *yeah*) are highlighted by the attention mechanism. These results also align with the intuition that the attention mechanisms are complementary.

4.4 Multi-view experiments

Our multi-view experiments use utterance-level representations to calculate the contrastive loss in

View1	View 2	Dev	Test	Comment
B-ACO-1	-	0.5858	0.5443	Acoustic-exclusive baseline
B-ACO-1	B-LEX	0.5971	-	Loss: $\mathcal{L}_c + \mathcal{L}_a + \mathcal{L}_m$ (Eqs. 2, 3, 4)
	H-MM-4	0.5976	-	Loss: $\mathcal{L}_c + \mathcal{L}_a + \mathcal{L}_m$ (Eqs. 2, 3, 4)
	H-MM-4 †	0.5969	-	Loss: $\mathcal{L}_c + \mathcal{L}_a$ (Eqs. 2, 3)
	H-MM-4 †	**0.6060**	**0.5859**	Loss: $\mathcal{L}_c + \mathcal{L}_{KL}$ (Eqs. 2, 5)
B-ACO-1 + Attention	H-MM-4 †	**0.6100**	**0.5976**	Loss: $\mathcal{L}_c + \mathcal{L}_{KL}$ (Eqs. 2, 5)

Table 3: The results of the multi-view experiments. We use the acoustic model B-ACO-1 as the first view and evaluate its performance using different second views. † means that the second view is not updated during training and its classification loss is not included.

Eq. 2. We discard experiments at the word level because 1) contrasting emotions for every word individually poses a complex task[2], and 2) context helps to disambiguate meaning as well as to convey the overall emotion rather than relying on high emotional words individually. Additionally, our experiments aim at a more practical scenario where there is no need for transcripts or ASR output with forced alignment.

Choosing negative samples. To calculate the loss as in Eq. 2, we randomly choose negative samples in two ways: 1) forcing a different class, and 2) forcing a different class that is acoustically similar to the positive sample (e.g., *sadness* vs. *neutral*, or *anger* vs. *happiness*). We saw that the model generalizes better using the second option. Our intuition is that the model does not have problems to force the margin m between vectors when the negative input samples come from fairly easy discriminative classes (e.g., *happiness* vs. *neutral*). In contrast, the model struggles to force the margin m between vectors when classes are acoustically similar, which turns into better generalization.

Different views. We choose B-ACO-1 as the first view because it uses raw frame level features. As shown in Table 3, we compare B-LEX and H-MM-4 as simple and elaborated second views by applying the contrastive and the views' cross-entropy loss functions. Indeed, by using B-LEX we show that the acoustic model B-ACO-1 improves its accuracy. Further improvements are made if we use H-MM-4 as a second view. This means that it is better to transfer information to the acoustic model when the modalities are effectively combined rather than when we try to induce only lexical information.

Frozen weights. We further explore H-MM-4 as a second view by first optimizing it, and then fixing its weights in the multi-view setting. Experiments with a trainable second view show that the lexical model is prioritized even when the losses are weighted as in Eq. 3 and 4. The intuition is that there is nothing new that this second view can learn from the multi-view setting once it has been optimized separately, and thus, it is better to exclude the complexity of learning it from scratch. Table 3 shows a small improvement over the previous models reaching 59.69% of UA on the validation set.

Teacher-student learning. We also experiment with a teacher-student setting where the model H-MM-4 is optimized separately. This model is a non-trainable second view where its class predictions are used as soft labels to evaluate the first view. The idea is to provide informative similitudes among the training samples by evaluating against a probability distribution over the classes rather than hard labels. The model reduces its loss more steadily than previous models, and once optimized, it surpasses previous results. Finally, we consider the case of a more complex student network since previous studies suggest that small student models may not be able to cope with the teacher models (Li et al., 2014; Meng et al., 2018). By adding an attention layer over the acoustic model B-ACO-1, we are able to improve the accuracy of the model by 1% absolute points, as shown in Table 3.

5 Conclusions

We presented multimodal and multi-view approaches for emotion recognition. The first ap-

[2] Negative words are hard to choose because we want properly formed utterances with the same number of words.

proach assumes that lexical information is always available when the speech signal is being processed. For such a scenario, our hierarchical multimodal model outperforms the state-of-the-art score with the aid of modality- and context-based attention mechanisms. The second approach adapts to a more realistic scenario where lexical data may not be available for evaluation. Our multi-view setting has shown that acoustic models can still benefit from lexical information over models that have been exclusively trained on acoustic features.

References

Zakaria Aldeneh, Soheil Khorram, Dimitrios Dimitriadis, and Emily Mower Provost. 2017. Pooling acoustic and lexical features for the prediction of valence. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI 2017, pages 68–72, New York, NY, USA. ACM.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–1247–III–1255. JMLR.org.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Adela Barbulescu, Rémi Ronfard, and Gérard Bailly. 2017. Which prosodic features contribute to the recognition of dramatic attitudes? *Speech Communication*, 95:78–86.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, COLT' 98, pages 92–100, New York, NY, USA. ACM.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Roddy Cowie. 2009. Perceiving emotion: towards a realistic understanding of the task. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1535):3515–3525.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 835–838, New York, NY, USA. ACM.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics.

Wanjia He, Weiran Wang, and Karen Livescu. 2016. Multi-view recurrent neural acoustic word embeddings. *CoRR*, abs/1611.04496.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167.

Qin Jin, Chengxin Li, Shizhe Chen, and Huimin Wu. 2015. Speech emotion recognition with acoustic and lexical features. 2015:4749–4753.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Jinyu Li, Rui Zhao, Jui-Ting Huang, and Yifan Gong. 2014. Learning small-size dnn with output-distribution-based criteria. In *Interspeech*.

Zheng Lian, Ya Li, Jianhua Tao, and Jian Huang. 2018. Speech emotion recognition via contrastive loss under siamese networks. In *Proceedings of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and First Multi-Modal Affective Computing of Large-Scale Multimedia Data*, ASMMC-MMAC'18, pages 21–26, New York, NY, USA. ACM.

N Majumder, D Hazarika, Alexander Gelbukh, Erik Cambria, and Soujanya Poria. 2018. Multimodal sentiment analysis using hierarchical fusion with context modeling.

Zhong Meng, Jinyu Li, Yifan Gong, and Biing-Hwang Juang. 2018. Adversarial teacher-student learning for unsupervised domain adaptation. pages 5949–5953.

Emily Mower Provost, Angeliki Metallinou, Chi-Chun Lee, Abe Kazemzadeh, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. 2009. Interpreting ambiguous emotional expressions.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2012. Understanding the exploding gradient problem. *CoRR*, abs/1211.5063.

Veronica Perez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 973–982, Sofia, Bulgaria. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Erik Cambria, Amir Hussain, and Alexander F. Gelbukh. 2018. Multimodal sentiment analysis: Addressing key issues and setting up baselines. *CoRR*, abs/1803.07427.

Viktor Rozgic, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, and Rohit Prasad. 2012. Ensemble of svm trees for multimodal emotion recognition. In *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific*, pages 1–4. IEEE.

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092.

Chang Xu, Dacheng Tao, and Chao Xu. 2013. A survey on multi-view learning. *CoRR*, abs/1304.5634.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *CoRR*, abs/1707.07250.

Appendix for "Multimodal and Multi-view Models for Emotion Recognition"

A Dataset Insights

This section describes some insights of the dataset. We use this information to take decisions relevant to our experiments. We consider the maximum number of words, the length of frames per utterance, and the number of frames per words.

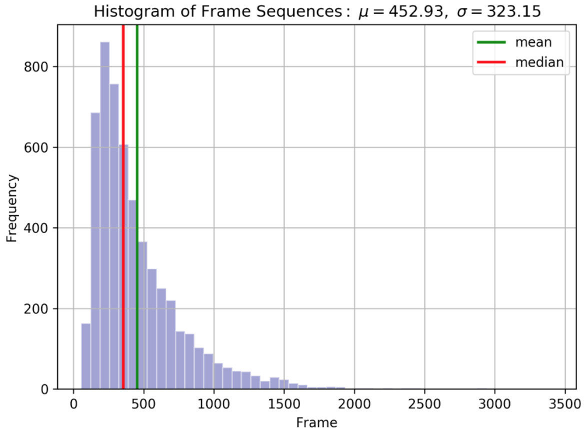

Figure 4: Histogram of frame sequences for every utterance.

We use 30 words as a maximum length for the sentences given that he average length is 17.40 and the standard deviation of 13.34 (see Figure 5). Additionally, we show statistics for the frame lengths on each utterance in Figure 4. We take a maximum length of 700 frames per utterance, where each frame is equivalent to 10 milliseconds.

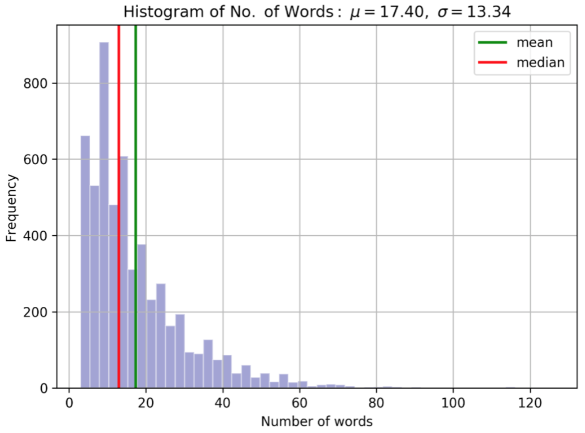

Figure 5: Histogram of number of words per utterance.

We also obtain the average length of frames that each word has according to the alignments of the dataset. Note that most of the words are within

100 frames, or equivalently, 1 second (see Figure 6).

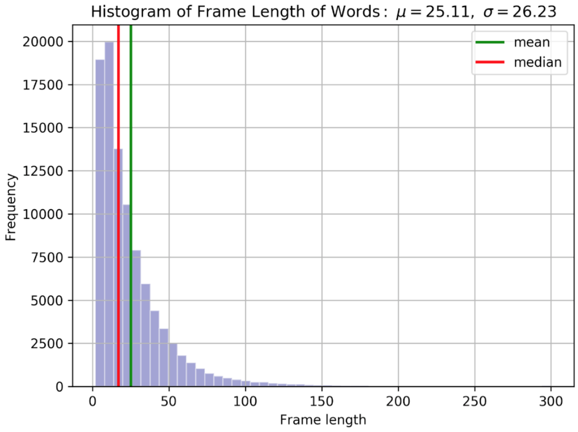

Figure 6: Histogram of word lengths in terms of frames.

B Experimental Settings

We train all our models for 30 epochs using a learning rate of 1e-4 and a batch size of 64. The optimization of the models is conducted using Adam (Kingma and Ba, 2014). We consistently use gradient clipping among our experiments. We clip the norm of the gradient beyond 5 (Pascanu et al., 2012; Goodfellow et al., 2016):

$$\mathbf{g} \leftarrow \frac{\mathbf{g}\tau}{||\mathbf{g}||} \text{ if } ||\mathbf{g}|| > \tau$$

To regularize the models, we use dropouts (Srivastava et al., 2014) by choosing drop probabilities between 0.4 and 0.5. We apply an ℓ_2 with a coefficient of 1e-5. For the GMU component, we use batch normalization applied to each modality matrix (Ioffe and Szegedy, 2015). All our experiments are validated using 10-fold cross-validation, leaving one speaker out of the training and validation sets.

For the multi-view learning experiments, we use the same settings as described for the multimodal experiments. In the case of the loss weights β_a and β_m, we experiment with values in $\{1.0, 1.2\}$ and $\{0.3, 0.5, 1.0\}$, respectively. We also experiment with βs as function of the epochs using

$$\beta = \frac{1}{1 + (\rho * epoch)}\beta_o$$

where ρ is a decreasing rate and β_o is the initial value, but the learning setting still overemphasize the multimodal view. The best results were achieved with $\beta_a = 1$ and $\beta_m = 0.3$ when both views were optimized simultaneously. For the margin in the contrastive loss function, we use $m = 0.5$.

For negative sampling in the contrastive loss function, we empirically found that using `anger` with `happiness` and `neutral` with `sadness` generally worked well since the acoustic patterns are similar. However, we saw some informative pairs when `happiness` and `anger` were coupled with `neutral`. This suggests that a more systematic way to determine pairs is needed. We leave the exploration of metrics such as valence, arousal and dominance to determine the contrastive pairs for future work.

C Additional Experiments

We run the following side experiments:

- Different length of words for our lexical baseline model (B-LEX). No benefit was perceived by going beyond 30 words.
- Different length of frames for our acoustic baseline model (B-ACO). The training time increases significantly while there is no substantial gain on performance by doing this.
- Improvised versus scripted utterances. We saw a substantial increase in performance (3%) of UA when speakers use scripted language rather than natural conversations.

D Model Insights

D.1 Visualization of Attention

We visualize the attention weights for correctly and incorrectly predicted emotions in Figures 7 and 8. Interestingly, when the sentences are read by humans, the target emotion for such utterances turn out ambiguous, which aligns with the result of the models.

D.2 Multi-view Results

By using the multi-view learning setting, we manage to induce lexical information into the model. According to Figures 9 and 10, it is easy to see that the model B-ACO-1 corrects a lot of the mistakenly predicted classes (i.e., compare neutral as ground-truth and sadness as prediction). However, the images also reveal that there are side effects such as transferring wrong aspects of the lexical modal to the acoustic one.

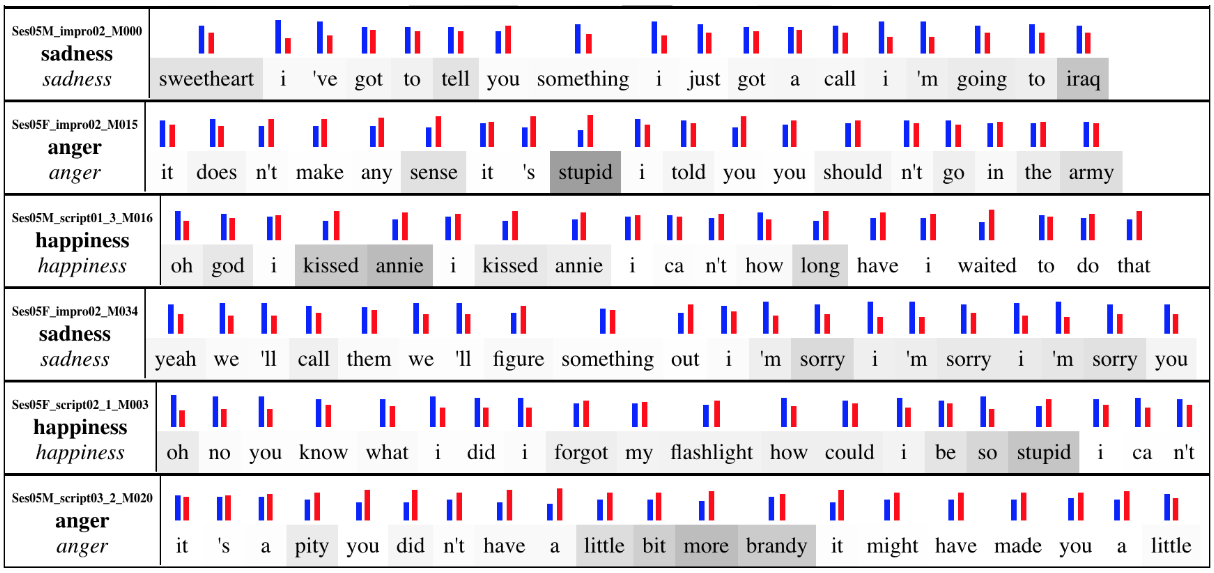

Figure 7: Correct predictions (italics) of the model along with the attention visualization.

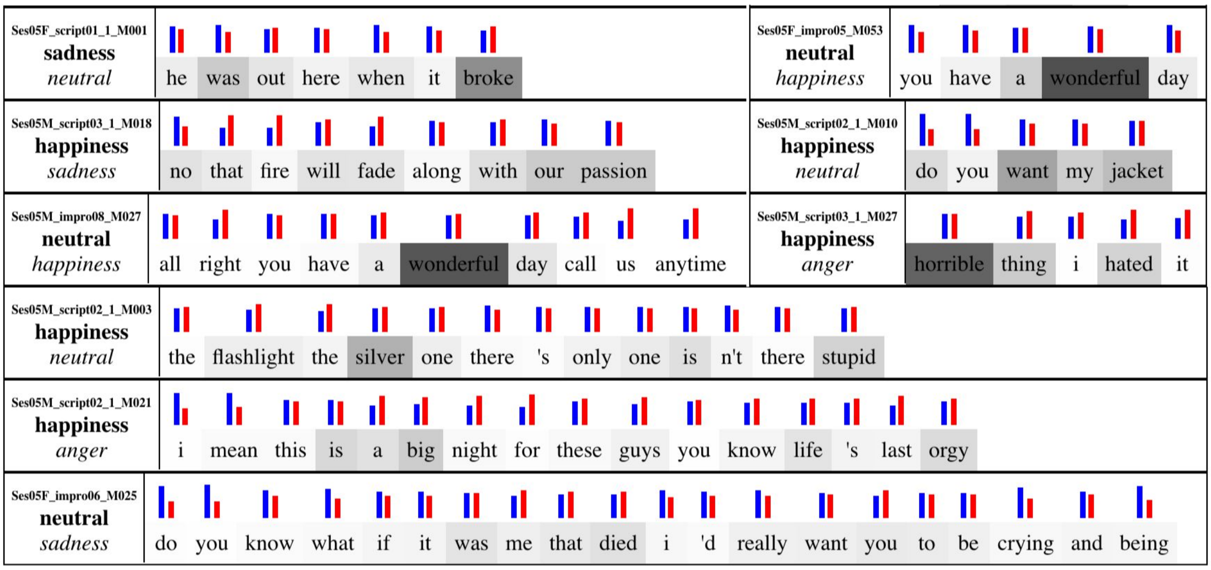

Figure 8: Incorrect predictions (italics) of the model along with the attention visualization.

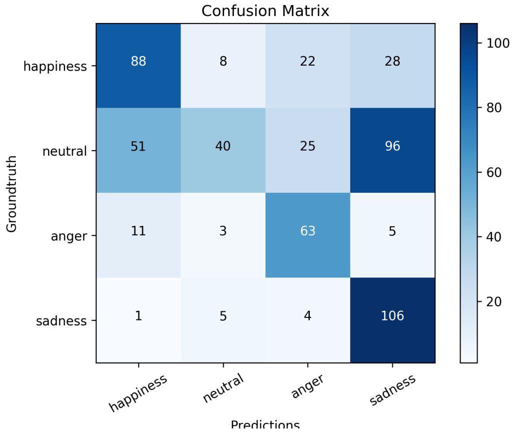

Figure 9: Confusion matrix of the acoustic model B-ACO-1.

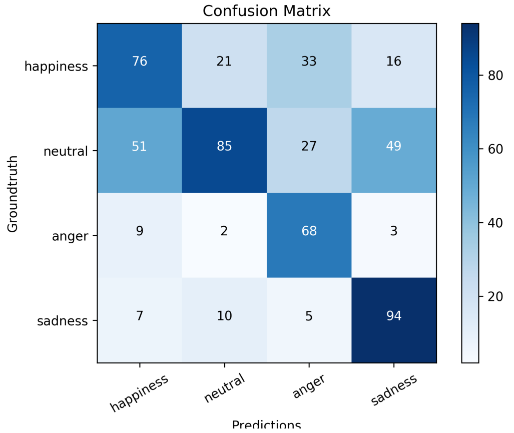

Figure 10: Confusion matrix of the acoustic model B-ACO-1 trained in a multi-view learning setting.