End-to-End Audiovisual Speech Recognition System with Multitask Learning

Fei Tao, *Student Member, IEEE,* Carlos Busso, *Senior Member, IEEE,*

***Abstract*—An *automatic speech recognition* (ASR) system is a key component in current speech-based systems. However, the surrounding acoustic noise can severely degrade the performance of an ASR system. An appealing solution to address this problem is to augment conventional audio-based ASR systems with visual features describing lip activity. This paper proposes a novel end-to-end, *multitask learning* (MTL), *audiovisual ASR* (AV-ASR) system. A key novelty of the approach is the use of MTL, where the primary task is AV-ASR, and the secondary task is *audiovisual voice activity detection* (AV-VAD). We obtain a robust and accurate audiovisual system that generalizes across conditions. By detecting segments with speech activity, the AV-ASR performance improves as its *connectionist temporal classification* (CTC) loss function can leverage from the AV-VAD alignment information. Furthermore, the end-to-end system learns from the raw audiovisual inputs a discriminative high-level representation for both speech tasks, providing the flexibility to mine information directly from the data. The proposed architecture considers the temporal dynamics within and across modalities, providing an appealing and practical fusion scheme. We evaluate the proposed approach on a large audiovisual corpus (over 60 hours), which contains different channel and environmental conditions, comparing the results with competitive *single task learning* (STL) and MTL baselines. Although our main goal is to improve the performance of our ASR task, the experimental results show that the proposed approach can achieve the best performance across all conditions for both speech tasks. In addition to state-of-the-art performance in AV-ASR, the proposed solution can also provide valuable information about speech activity, solving two of the most important tasks in speech-based applications.**

***Index Terms*—audiovisual speech recognition, deep learning, multitask learning, end-to-end speech systems**

I. Introduction

SPEECH processing technology that enables hand-free interaction with systems have revolutionized commercial applications (e.g., Google Voice Assistant, Alexa and Siri). Among speech processing systems, *voice activity detection* (VAD) and *automatic speech recognition* (ASR) are crucial components to understand when the user is speaking, and what the user is saying. While current ASR technology has reached human-level performance in clean conditions [1], the performance often drops in real world conditions with a noisy acoustic environment. Introducing visual cues is an appealing approach to improve the robustness of ASR systems [2]–[4]. With advances in multimedia technology, studies across areas have consistently demonstrated the improvement in performance by fusing multimodal information [5]–[10]. These studies have motivated the community to introduce visual cues for more robust ASR systems [3], [4], [11].

Conventional *audiovisual automatic speech recognition* (AV-ASR) relies on hand-crafted visual features [12]. Recently, variations of *deep neural networks* (DNN) have emerged as powerful alternatives for feature space modeling, which have made end-to-end frameworks feasible [13]. An end-to-end framework directly learns discriminative characteristics from the raw input, extracting high-level feature representation to predict the labels of the given task. The end-to-end approaches are appealing as they do not need to manually define rules, which may be too rigid and constrain the systems' performance. They can learn all the required information directly from the data. For ASR, applying the *connectionist temporal classification* (CTC) [14] loss function can provide the capability of temporal classification on unsegmented sequence for DNN. While the DNN framework trained with the CTC loss function can be extended to AV-ASR tasks [15], the original CTC does not have timing alignment information, which can limit the model capability. Having timing information about speech activity can help an ASR system to make more reliable predictions. This observation suggests that combining AV-ASR and *audiovisual voice activity detection* (AV-VAD) tasks can be beneficial to increase the model capability of the speech-based system.

This study proposes a *multitask learning* (MTL) technique that combines AV-ASR and AV-VAD tasks. MTL explores the advantages of solving multiple related tasks with a unified framework [16]. By jointly solving multiple tasks, the network learns feature representations that are discriminative for all tasks, creating solutions with better generalization than models trained using *single task learning* (STL). MTL learning is an appealing approach in speech processing [17]–[24], since similar acoustic features conveying rich spectral information are often used for various tasks including speaker verification, ASR, and speech emotion recognition. AV-ASR is the primary task in our formulation, since this is the focus of our study. The secondary task is AV-VAD. The AV-VAD task identifies segments with speech activity providing valuable timing information leveraged by the CTC loss function, leading to improvements in AV-ASR.

The proposed MTL structure has separate sub-networks for acoustic and visual features, which are later combined with *recurrent neural network* (RNN). The acoustic features correspond to Mel-scaled filterbank features. The visual features are directly extracted from the pixels with *convolutional neural networks* (CNNs), which can learn informative high-level rep-

F. Tao was with the Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson, TX, 75080 USA e-mail: fei.tao@utdallas.edu.

C. Busso is with the Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson, TX, 75080 USA e-mail: busso@utdallas.edu.

Manuscript received March 28, 2019; revised xxx, 2019.

resentation from raw input [25]. After combining the unimodal sub-networks, we implement separate task-dependent *fully connected* (FC) layers for both AV-ASR and AV-VAD. The global system combines the loss function of the AV-ASR task (CTC), with the loss function of the AV-VAD task (cross-entropy). The proposed approach is evaluated with a subset of the CRSS-4ENGLISH-14 corpus [26], consisting of over 60 hours of audiovisual recordings. The results show clear improvements over other competitive audiovisual baselines, not only for the primary task, but also for the secondary task. The experimental results also demonstrate that the proposed end-to-end approach to combine audiovisual information is effective in increasing the robustness for ASR tasks. This result is demonstrated with evaluations on the GRID corpus.

The remaining part of this paper is organized as follows. Section II describes relevant studies related to our work. Section III describes the corpus used for this study and the audiovisual features. Section IV introduces our proposed approach, presenting the details of the framework. Section V discusses the implementation of our approach and the baselines used to compare our method. Section VI presents the experiments performed to evaluate the proposed approach, comparing the results with competitive baselines. Section VII concludes the paper, presenting future research directions.

II. Related Work

Previous studies have shown that audiovisual solutions can improve the robustness of speech processing systems when these systems are tested under mismatched speech mode (e.g., whisper speech) or acoustic noisy environment [2], [27]–[32]. The interest in this area has increased with recent advances on audiovisual corpora for continuous speech recognition [26], [33], [34]. Conventional audiovisual solutions rely on manually defined rules to fuse the modalities, using handcrafted features [35], [36]. These systems tend to lack flexibility, limiting the modeling capability of the algorithms. If the models are end-to-end, these systems can also automatically learn discriminative features directly from the data.

One of the first studies on audiovisual fusion using DNN was presented by Ngiam et al. [37]. They proposed several configurations for AV-ASR to fuse audiovisual features, including a bimodal neural network using an autoencoder. Tao and Busso [26] proposed a *gating neural network* (GNN) for AV-ASR, which filtered noisy information in the presence of mismatched conditions. The gating layers worked as a switch, allowing or denying information going through the neurons. The parameters of the gating layers were directly learned from the data. Recent studies have used *recurrent neural network* (RNN) with attention model to improve the temporal modeling information, increasing the flexibility of the audiovisual models [38], [39]. Studies have also proposed fusion schemes for AV-VAD [35], [36], [40]–[43]. Using DNN, Tao and Busso [44] proposed a *bimodal recurrent neural network* (BRNN) to fuse the temporal information across modalities. The BRNN architecture creates sub-networks for the individual modalities, which learn their characteristic temporal dynamics. The sub-networks are then fused with *bidirectional long short-term memory* (BLSTM) layers, capturing the temporal relationships across modalities. While all these studies have increased the flexibility of the systems to fuse modalities, they considered handcrafted audiovisual features.

Studies have extended DNN techniques to extract discriminative features directly from the raw data during training. Petridis et al. [45] trained an autoencoder to extract visual features for lip reading. The autoencoder reconstructs the lip area by passing the information over a bottleneck layer with reduced dimension. The activations of this bottleneck layer are then used as visual features. Stafylakis and Tzimiropoulos [46] used *residual network* (ResNet) with CNNs to extract visual features. CNNs learn transform-invariant visual representations, which is appealing for this application. The use of ResNet was useful to learn the input feature space relying on a deep architecture.

End-to-end systems for speech processing tasks have clear potential, since they are more flexible than conventional approaches, without making rigid constraints that are often made with manually designed rules. Using only speech, Adomei et al. [13] and Graves and Jaitly [47] proposed end-to-end DNN systems, learning high-level acoustic representations from raw audio input. These studies have shown that end-to-end frameworks can reach better performance than alternative methods, since (1) they extract high-level representation directly from raw inputs, without discarding information by preprocessing the data, (2) they jointly learn all the parameters, so achieving global optimal configuration is possible. Given these advantages, studies have proposed end-to-end audiovisual systems. For AV-ASR, Petridis et al. [48] proposed an end-to-end system based on ResNet and *bidirectional gated recurrent units* (BGRUs). The end-to-end AV-ASR system led to improvements in within-context word recognition over audio-only systems, especially in the presence of acoustic noise. For AV-VAD, Tao and Busso et al. [44], [49] implemented the BRNN framework in an end-to-end fashion, combining feature space modeling, temporal modeling (inter and intra modalities), audiovisual fusion and classification into one neural network. The system outperforms the state-of-the-art baselines, indicating the capability of the end-to-end framework. The audio inputs in the aforementioned end-to-end systems relied on audio features that were barely processed by a front-end before entering the network as features (e.g,. spectrogram, Mel-scaled filterbank). We follow the same naming convention in this paper to refer to our framework as end-to-end system (we use Mel-scaled filterbank as our audio input).

Graves et al. [14] proposed the CTC loss function to train their end-to-end ASR framework. The CTC loss function can learn the mapping between acoustic features and sequence of unsegmented labels (e.g., sequence of characters). The input and output sequences can have different lengths. This approach is suitable for ASR tasks, where input (features) and output (characters) sequences have different lengths. We hypothesize that adding timing information about speech segments can help an AV-ASR system trained with a CTC loss function, providing complementary timing information. However, exact information about segments is difficult to obtain (e.g., phoneme label per frame). Recently, Petridis et

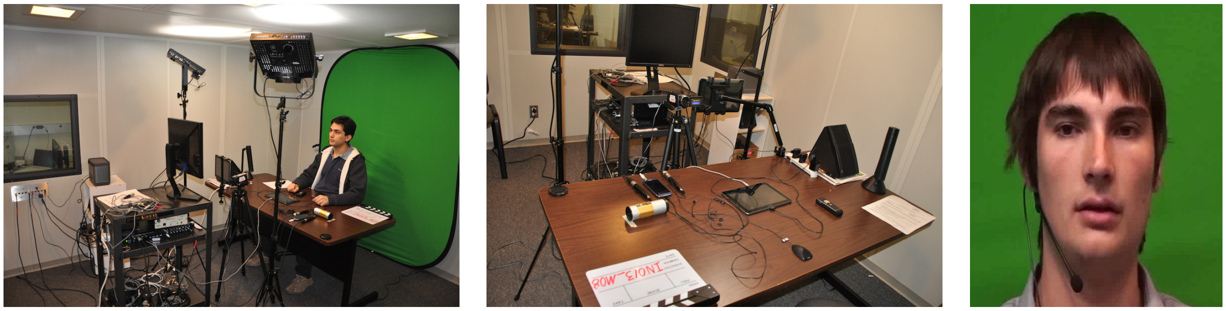

(a) (b) (c)

Fig. 1. Data collection for the CRSS-4ENGLISH-14 corpus. (a) sound booth used to record the data, (b) experimental setup with the sensors for the data collection, (c) a frame displaying the face of one of our subjects.

al. [50] proposed the use of attention model as an auxiliary task with CTC to improve the performance of an AV-ASR system. The attention model is expected to learn the temporal alignment information. However, learning the parameters of an attention model is computationally expensive. As an alternative, information about segments with speech activity is easier to obtain, given the recent advances in AV-VAD. VAD systems have been successfully used to improve the performance of other speech tasks. Motivated by these observations, this study proposes a MTL framework that uses an AV-VAD task as a secondary task, while training the AV-ASR system (e.g., primary task). With the text labels and timing information, the proposed MTL framework can help to better mine the input signals, learning a more discriminative and robust high-level feature representation [51]. This is a novel end-to-end MTL framework for AV-ASR that jointly learns discriminative features and coupling relationships between the modalities in a unified and systematic fashion.

III. Data Preparation

A. Corpus Description

This study uses the CRSS-4ENGLISH-14 corpus [26], which was collected in a $13ft \times 13ft$ sound booth (Fig. 1(a)). It consisted of 442 subjects speaking English, in four different accents: American (115), Australian (103), Indian (112) and Hispanic (112). We only used the data from American speakers, which have a duration of 60 hours and 48 minutes. The CRSS-4ENGLISH-14 corpus was collected with multiple cameras and microphones (Fig. 1(b)). This study only considers audio from a close-talking microphone (Shure Beta 53) and a tablet microphone (Samsung Galaxy Tab 10.1N), and video from a *high definition* (HD) camera (Sony HDR-XR100) and a camera from a tablet (same tablet used to record the audio). With these sensors, we set two conditions. The *ideal channel* consists of audio from the close-talking microphone and video from the HD camera, which provide high quality data from the speakers. The *challenging channel* consists of audio and video from the tablet, which was set about two meters from the speakers. Figure 1 describes the experimental setup of the data collection, including an example of a frame displaying the face of one of our subjects (Fig. 1(c)).

The database includes read and spontaneous speech. In the first part of the data collection, the speakers read scripts including sequence of digits, city names, short commands, continuous sentences, and isolated words. They also answered prompted questions (spontaneous speech). We refer to this part as *clean recordings*. In the second part, we introduced noise with a loudspeaker (Beolit 12), playing five types of pre-recorded noises: in-car, home, restaurant, shopping mall and office. The loudspeaker was placed closer to the tablet, so the impact of the noise is stronger in the *challenging channel*. For the second part, we randomly included a subset of read recordings used in the first part. We refer to this part as *noisy recordings*. We manually transcribed the collected data. Tao and Busso [26] provides a more detailed description of the corpus.

B. Feature Extraction

The proposed model takes raw audiovisual features as inputs. For the visual input, we extract gray scale images around the mouth area. We use the toolkit IntraFace [52] to detect 49 facial landmarks on each frame. We observed that the tool failed to detect the facial landmarks on 0.1% of the frames. We discard videos if more than 20% of their frames are not properly processed by IntraFace. For the rest of the videos, we use linear interpolation to recover the missing frames. One frame with frontal pose is manually selected as a global template. We normalize the face with an affine transformation that compares rigid landmarks between each frame and the frontal template. The rigid points correspond to facial landmarks that barely move across facial expressions (e.g., points around the nose). The resulting images have similar size and pose after the normalization. From this normalized image, we extract the *region of interest* (ROI) around the mouth area, where the center is located at the centroid of the lip landmarks, and its size is downsampled to 32×32 pixels. For the audio input, we extract 26D Mel-scaled filterbank, relying on the tool *python speech features* [53].

IV. Proposed Approach

Figure 2(a) describes the proposed end-to-end MTL bimodal RNN architecture. The framework has two unimodal sub-networks that separately process the audio and visual inputs. These sub-networks are referred to as *audio RNN* (A-RNN) and *visual RNN* (V-RNN). The third sub-network takes as input the concatenated hidden values from the A-RNN and V-RNN sub-networks, fusing the modalities. We refer to this sub-network as *audiovisual RNN* (AV-RNN). Formulating this task as a multitask learning problem, there are two classification layers on top of the AV-RNN network, where the primary goal is AV-ASR and the secondary task is AV-VAD. While we presented some of the blocks of this framework in our previous work [44], [49], [54], the use of multitask learning is a novel contribution of this study. From an implementation perspective, the proposed approach is practical as it does not need extra annotations. From the transcription needed for the ASR task and using forced alignment, it is straightforward to derive labels for the VAD task. Therefore, our novel solution improves the ASR performance without any extra cost. This section presents the details of each component of the proposed approach.

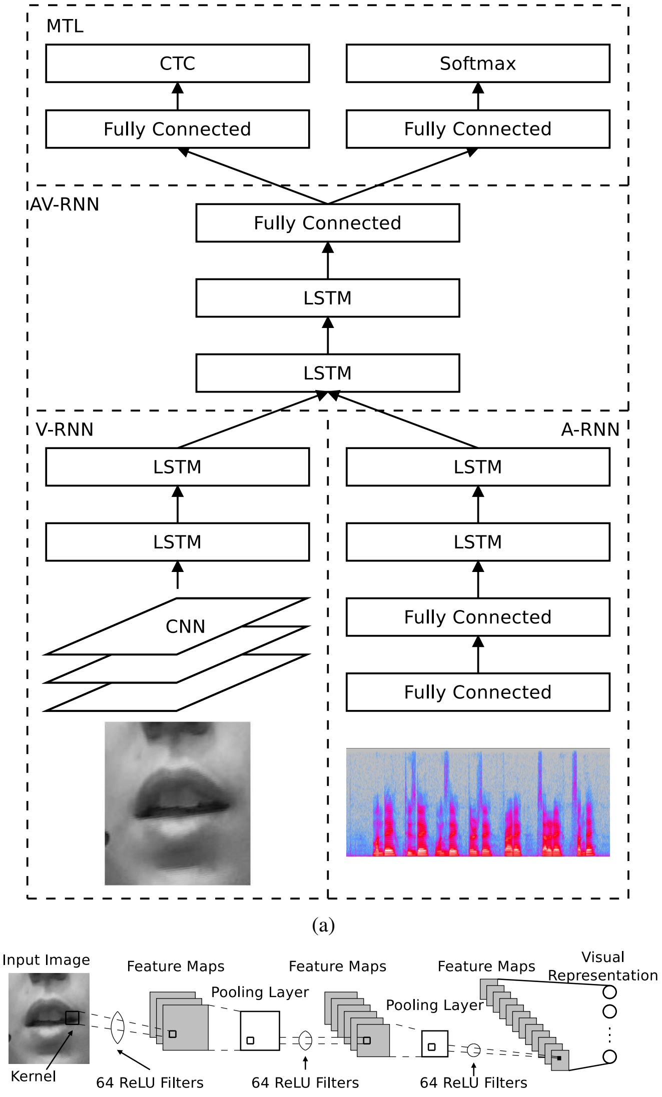
A. Unimodal Sub-networks

The A-RNN sub-network takes Mel-scaled filterbank feature as input. The network processes the input signals relying on two FC layers with 256 *rectified linear units* (ReLUs). The objective from these layers is to obtain discriminative representation from the raw input. We did not use convolutional layers to process the audio to reduce the computing resources to train the model. On top of the FC layers, we use two unidirectional *long short-term memory* (LSTM) layers with 256 neurons. The LSTM layers capture the temporal dependency within the audio modality, improving the feature representation. The use of unidirectional LSTM instead of BLSTM leads to low latency, which is important for real-time implementation.

The V-RNN sub-network takes raw pixels from the video frames as the visual input and learns a high-level feature representation relying on CNN. The CNN captures the mouth appearance and shape information relevant to the target task with learnable filters, which can be trained with back propagation. Figure 2(b) shows the proposed structure. We use pooling layers after the convolution layers to make the visual representation invariant to spatial transformation. This study uses three convolutional layers, each of them implemented with 64 ReLU filters. The first convolutional layer has a kernel size of 5×5 and stride size of two. A max pooling layer is applied after the first convolutional layer. The pooling size is two. The second convolutional layer has a kernel size of 3×3 and stride size of two. A max pooling layer with pooling size two is applied after the second convolutional layer. The third convolutional layer has a kernel size of 3×3 and stride size of one. After the third convolutional layer, hidden values are flattened and used as input of two unidirectional LSTM layers with 64 neurons per layer. The LSTM layers further process the extracted visual representation and model the temporal dependency within the visual modality.

B. AV-RNN

The hidden values from the top LSTM layers in the A-RNN and V-RNN sub-networks are concatenated and used as input for the AV-RNN sub-network. The AV-RNN sub-network fulfills three important roles: 1) it fuses the two modalities; 2) it models the temporal dependency across modalities; 3) it further processes the feature representation for the classification tasks. The AV-RNN directly learns the weights associated with the modalities during training, which is an advantage over previous formulations relying on hyper-parameters. The temporal dependencies across modalities are captured with two unidirectional LSTM layers (256 neurons per layer). Finally, we add a FC layer with 256 neurons implemented with ReLUs.



(a)

(b)

Fig. 2. (a) Diagram of the proposed end-to-end MTL framework, which consists of three sub-networks. The A-RNN sub-network processes raw audio input relying on FC and LSTM layers. The V-RNN sub-network processes raw video input relying on CNN and LSTM layers. The activation from the A-RNN and V-RNN sub-networks are concatenated and sent to the AV-RNN sub-network. The final high-level feature representation is used to solve two speech tasks: AV-ASR and AV-VAD. (b) Diagram of the CNN-based architecture used in the V-RNN sub-network.

C. Multitask Layers

The main contribution of this study is the use of MTL for audiovisual speech recognition. On top of the AV-RNN, we simultaneously address speech recognition (AV-ASR), which is the primary task, and speech activity detection (AV-VAD), which is the secondary task. AV-VAD and AV-ASR are related tasks. For example, without speech activity, the AV-ASR should not produce an output. Therefore, the representation containing timing information from the AV-VAD is expected to be helpful for the AV-ASR task. Each task consists of one FC layer followed by one classification layer (Fig. 2(a)). For the AV-ASR task, the FC layer has 256 ReLU neurons and the classification layer is a softmax layer using CTC as the loss function. For the AV-VAD task, the FC layer has 256 ReLU neurons and the classification layer is a softmax layer using cross-entropy as the loss function. The loss function $\mathcal{L}$ is defined with Equation 1, where y is the ground truth of the VAD task, $\hat{y}$ is the prediction of the VAD task, π is all possible paths for the ASR task, B is the entire ASR search space, x is the input sequence and θ are the network's learnable parameters (e.g., weights and bias).

$$\begin{aligned}\mathcal{L} = \arg\min_{\theta}(-\alpha_{asr} \sum_{\pi \in B(x,y)} P(\pi|x)) \\ +\alpha_{vad}[y \ln \hat{y} + (1-y)\ln(1-\hat{y})] \qquad (1)\end{aligned}$$

The term with the summation in Equation 1 describes the CTC loss, which is computed for the entire sequence. Equation 2 further explains the CTC loss. The entire sequence π consists of labels l_t from time 1 to T. The labels in this study are characters. The total conditional probability is the product of the posterior probability across all the time steps. $P(l_t, t|x)$ is the posterior probability of the output of the neural network shown in Equation 3, which is the result of the softmax function and can be trained at each time step by error back propagation with the cross-entropy loss. The negative sign in Equation 1 is introduced to maximize the posterior probability over the entire sequence.

$$P(\pi|x) = \prod_{t=1}^{T} P(l_t, t|x) \qquad (2)$$

$$P(l|x) = Softmax(x) \qquad (3)$$

The term $y \ln \hat{y} + (1-y)\ln(1-\hat{y})$ in Equation 1 corresponds to the cross-entropy loss for the AV-VAD task, which is computed for each frame. This is a binary classification task where the labels indicate the presence or absence of speech. This term regularizes the network increasing the generalization and robustness of the AV-ASR task. The parameters α_{asr} and α_{vad} in Equation 1 are hyper-parameters representing the weights of the AV-ASR task and AV-VAD task for training the multitask learning system. We set the weights for both tasks with the same value, acknowledging that even better performance than the ones reported in this study can be achieved by optimizing these hyper-parameters (Section VI-D explores the sensitivity of the weights by comparing this model with alternative settings, where we weight more either the AV-ASR loss or the AV-VAD loss).

The end-to-end system is jointly trained sharing the timing and character information to intermediate layers, deriving powerful representations that are discriminative for both tasks. By jointly solving both problems, we simultaneously improve the performance of both tasks, as demonstrated in the experimental evaluation.

V. Experiment Settings

A. Implementation of the Multitask AV-ASR

We evaluate the proposed approach on the CRSS-4ENGLISH-14 corpus. We have 10 speakers from the US group, where we had problem with the videos. We discarded the data from these subjects, resulting in 105 speakers for the experiments. The train set uses recordings from 70 speakers, the test set uses recordings from 25 speakers, and the validation set uses recordings from 10 speakers. All these partitions are gender balanced. We test the model using four conditions: *ideal channel* (close-talking microphone and HD camera) under clean and noisy environment, and *challenging channel* (microphone and camera from tablet) under clean and noisy environment (Section III-A). In contrast, we train the system only with the *ideal condition* using clean recordings. Our rational for not using noisy data during training is that the mismatch between train and test conditions is always present in practical applications. The goal of our work is to develop a well-generalized and robust AV-ASR system, which can reduce the mismatch as much as possible. This principle led us to fix the training data and focus on the models, which aim to improve the performance under different conditions as much as possible.

To capture more temporal information, we concatenate the current acoustic frame with 10 previous frames creating a contextual feature vector of 11 frames. For the visual input, we only use the current frame without concatenating previous frames, simplifying the complexity and memory requirements in training the system. For the AV-ASR task, we use the sequence of characters consisting of 37 classes (26 English letters, 10 digits and a blank class). For the AV-VAD task, we use the tool SAILAlign [55] for force alignment, using the manually transcribed recordings. The toolkit did not provide an alignment on 2.1% of the segments, which are discarded in this study. After alignment, we create a binary label per frame to indicate the presence or absence of speech (speech versus non-speech classes). We use the *character error rate* (CER) as the performance metric for the AV-ASR task. Since we focus on the performance of the acoustic model, we do not develop a language model, which is often used to report the conventional *word error rate* (WER). We perform the McNemar's test to assert whether the differences in ASR performance between two systems are statistically significant at p-value=0.01. While the focus of our research is on AV-ASR (i.e., primary task), we also obtain valuable information from AV-VAD (i.e., secondary task) as a side benefit of the proposed multitask framework. Therefore, we also report the performance of the AV-VAD system when it is applicable. We use the F-score as the performance metric for the secondary task, where speech is the target class.

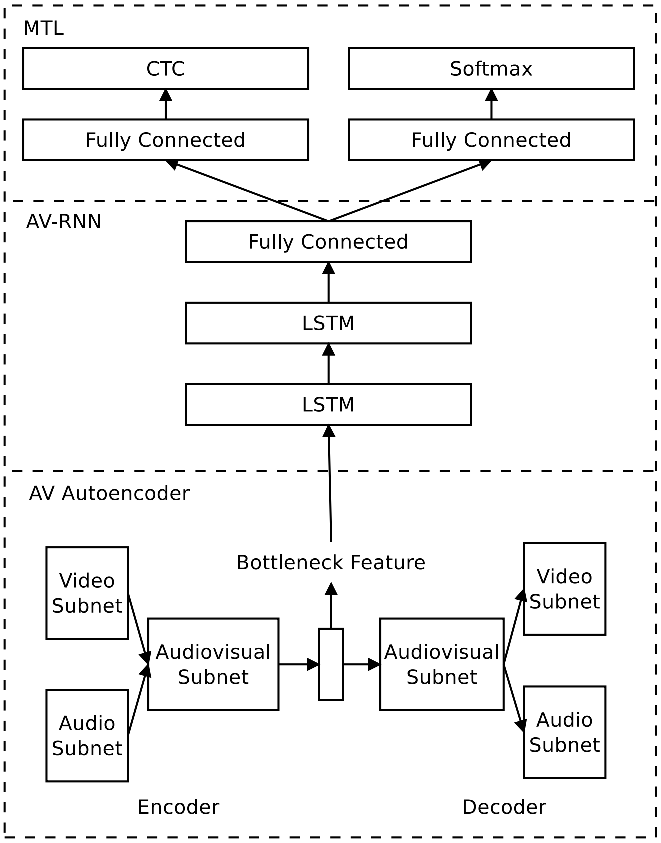


Fig. 3. Diagram of the “AE-RNN-MTL” baseline. The approach consists of two systems, which are separately trained. The first system is the audiovisual autoencoder, which is used to extract the bottleneck feature. The second system replicates the AV-RNN sub-network and MTL layers of our proposed approach. The “AE-RNN-STL” baseline has the same architecture as the “AE-RNN-MTL” framework, with the exception that the loss function only considers the AV-ASR task, ignoring the AV-VAD task.

The framework is trained with standard techniques. We use ADAM optimizer [56] to train all the systems, using a dropout rate [57] set to 0.1. We monitor the loss on the validation set. We rely on early stopping when the validation loss decreases less than 5% within three consecutive epochs. The experiments are run on a Nvidia GTX 1070 graphic card with 8 GB memory.

B. Baseline Methods

We implement three baselines to compare our proposed approach. Since our main focus is AV-ASR, the baselines are implemented to solve this problem. The first baseline, shown in Figure 3, is inspired by the classic AV-ASR system proposed by Ngiam et al. [37], which uses an autoencoder. This approach is a competitive baseline for audiovisual tasks. While this approach was originally presented in 2011, several current unimodal and multimodal systems continue to use variations of this model given the competitive performance achieved by this framework. We implemented a bimodal autoencoder system to fuse audiovisual features. The encoder consists of three sub-networks: an audio sub-network, a visual sub-network, and an audiovisual sub-network. The visual sub-network is the same as the convolutional structure in the V-RNN sub-network in the proposed network (Figure 2(b)). The audio sub-network has three FC layers with 256 ReLUs per layer. The hidden values from the top layers of the audio and visual sub-networks are concatenated and sent to the audiovisual sub-network. From bottom to top, the audiovisual sub-network has three layers with 256, 128 and 64 ReLU activations. On top of these layers, we implement a bottleneck layer with 32 ReLU neurons. The architecture of the decoder is a reversed mirror of the encoder. It starts from the bottleneck layer and it ends with the visual and audio sub-networks. The activations of the bottleneck layer are used as the fused audiovisual features that are fed as input of a RNN to recognize speech (Fig. 3). The AV-RNN network has two unidirectional LSTM layers with 128 ReLU neurons per layer, and one FC layer with 128 ReLU units. The MTL network predicts the ASR and VAD tasks. Each task has a FC layer with 128 ReLU neurons. The output layer for the ASR has 37 classes, and the output layer for the VAD task has two classes. This baseline is referred to as “AE-RNN-MTL”. For the second baseline, we implement this autoencoder approach as a single task learning problem, where the loss function consists of only the CTC loss. We refer to this system as “AE-RNN-STL”. It should be noticed that the bottleneck representation using the encoder is fixed and used to train the AV-ASR (and AV-VAD) framework. Therefore, the features are not extracted to optimize the performance of the supervised tasks as in our multitask BRNN framework. This key different allows us to directly compare the benefit of learning the feature representation directly from the raw data in one pass (proposed end-to-end approach) and learning the feature representation as a separate step (two-step approach in AE-RNN-STL and AE-RNN-MTL).

The third baseline system compares our multitask learning approach with an equivalent system optimized to only recognize speech (*single-task learning* (STL)). The baseline system is set with the same architecture as the proposed approach. The only difference is the cost function, which only considers the AV-ASR task. We refer to his baseline as “STL”. The STL baseline is meaningful since it allows us to directly compare the addition of the secondary tasks, which is a major contribution on this study.

VI. Experimental Results

A. AV-ASR Results

Table I shows the experimental results. The first part of the table lists the results when the models are tested with the ideal channel. Under clean environment, the proposed end-to-end MTL system can significantly outperform the baseline systems (p-value<0.01). It is better than the “STL” system by 6.4% (absolute). The only difference between both conditions is the cost function which also considers the AV-VAD task while training the MTL system. The proposed system is also better than the “AE-RNN-MTL” and “AE-RNN-STL” systems by 22.7% and 24.5%, respectively (absolute), which highlights the importance of directly learning a high-level representation within a single framework directly from the raw data. For the autoencoder frameworks, the features are not extracted to optimize the performance of the supervised

TABLE I
Experiment results for various environment and channel conditions. "MTL" refers to the proposed multitask learning system, "STL" refers to the single-task learning system with only the AV-ASR task. The baselines "AE-RNN-MTL" and "AE-RNN-STL" refer to the systems inspired by the work of Ngiam et al. [37], implemented with MTL and STL, respectively. The table reports CER for ASR, and F-score for VAD (ideal: close-taking microphone, HD camera; challenging: microphone and camera from tablet). Font in bold shows significant improvement (p-value<0.01).

Channel	Env.	Model	AV-ASR	AV-VAD
			CER [%]	F-score [%]
Ideal	Clean	MTL	**12.5**	95.3
		STL	18.9	-
		AE-RNN-MTL	35.2	94.6
		AE-RNN-STL	37.0	-
	Noisy	MTL	**15.0**	95.5
		STL	22.5	-
		AE-RNN-MTL	39.8	95.0
		AE-RNN-STL	41.6	-
Challenging	Clean	MTL	**15.8**	94.7
		STL	19.5	-
		AE-RNN-MTL	35.6	94.1
		AE-RNN-STL	37.8	-
	Noisy	MTL	**36.5**	91.9
		STL	36.9	-
		AE-RNN-MTL	51.7	91.5
		AE-RNN-STL	52.2	-

task. In contrast, our proposed method is end-to-end, where the feature representation is directly learned by minimizing the loss function of the supervised tasks. This is a key difference, which leads to clear benefits in using the proposed approach.

For the data from the ideal channel under noisy environment, the results are similar. The proposed approach can significantly outperform the baseline systems by a large margin (p-value<0.01). The absolute improvements compared with the "STL", "AE-RNN-MTL" and "AE-RNN-STL" systems are 7.5%, 24.8% and 26.6%, respectively. The improvement shows the benefit of the proposed approach even in noisy acoustic environment.

The second part of Table I shows the performance when tested with the challenging channel. Under clean environment, the proposed approach can significantly outperform the baseline systems (p-value<0.01). The absolute improvements compared with the "STL", "AE-RNN-MTL" and "AE-RNN-STL" systems are 3.7%, 19.8% and 22.0%, respectively. Even when we have a channel mismatch in training and testing the models, the proposed MTL approach still achieves the best performance.

The most difficult setting is when we have a channel mismatch with noisy recordings (challenging channel with noisy speech). For this condition, the performance of the MTL framework is still better than the baselines. It significantly outperforms the "STL", "AE-RNN-MTL", and "AE-RNN-STL" baselines by 0.4%, 15.2%, 15.7%, respectively (p-value<0.01). We notice that the improvement over the "STL" baseline is limited (0.4%), compared to other testing conditions. We hypothesize that the information inferred by the AV-VAD is not accurate enough due to the noisy environment, which also affects the AV-ASR task. The noise used in our setting was collected from malls, car, restaurants, offices, and home environments, which often includes human voice in the background. This type of noise is challenging for VAD, especially when the mismatch between train and test conditions increases. For all other conditions we obtain important benefits by using MTL over STL.

Another powerful result observed in Table I is that the MTL version of the baseline (AE-RNN-MTL) achieves better performance than its implementation with STL (AE-RNN-STL). This result confirms our hypothesis that adding timing information from the AV-VAD task (secondary task) to the AV-ASR task trained with CTC loss leads to better performance and robustness. This is an important finding in this study.

B. AV-VAD Results

While the primary task is AV-ASR, Table I also shows the performance for the AV-VAD system, which is our secondary task. Using the ideal channel under clear condition, the proposed system achieves an F-score of 95.3%. In Tao and Busso [54], we proposed a single task learning, end-to-end AV-VAD system. Under similar conditions, this framework achieved an F-score of 94.0%. The comparison shows that using MTL can also improve the secondary task performance, even when the primary task is AV-ASR.

For the challenging channel under clean condition, the performance for the AV-VAD task (secondary task) is 1% better than the results of our single task learning, end-to-end AV-VAD system reported in Tao and Busso [54] (F-score of 93.7% for this setting). The performances of the MTL framework for the AV-ASR and AV-VAD tasks show that this structure can work well with the data collected with the tablet.

C. Performance as a Function of Types of Noise

We also present the results of our approach as a function of the type of noise. This analysis considers the challenging channel under noisy condition. Table II shows the results for each type of noise. The table shows that the office environment has the best performance with a CER of 29.2% for the AV-ASR task, and a F-score of 93.6% for the AV-VAD task. The worst performance was observed with noise collected in a shopping mall (CER of 52.0% for AV-ASR; F-score of 90.0% for AV-VAD). The office environment is characterized by keyboard noise and occasional human talk, so this type of noise is not so detrimental to the speech tasks. In contrast, the noise in the mall contains human speech at all time, creating a challenging condition. Noise recorded in restaurants and in cars are also challenging for our speech tasks.

D. Weights of Primary and Secondary Tasks

The results reported in previous sections consider a MTL system trained with equal weights for the primary (AV-ASR) and secondary (AV-VAD) tasks ($\alpha_{asr} = 1$ and $\alpha_{vad} = 1$ in Equation 1). Ideally, we should use the validation set to finding optimal values for the weights. However, training the model is computationally intensive, so this option is not feasible with our current infrastructure. Instead, we explore the sensibility of the results by evaluating two alternatives configurations,

TABLE II
PERFORMANCE OF THE PROPOSED APPROACH AS A FUNCTION OF THE TYPE OF NOISES. THE EVALUATION CONSIDERS THE CHALLENGING CHANNEL WITH NOISY CONDITION.

Noise Type	AV-ASR CER [%]	AV-VAD F-score [%]
In-car	35.5	91.9
Home	32.8	92.6
Restaurant	36.8	91.2
Mall	52.0	90.0
Office	29.2	93.6

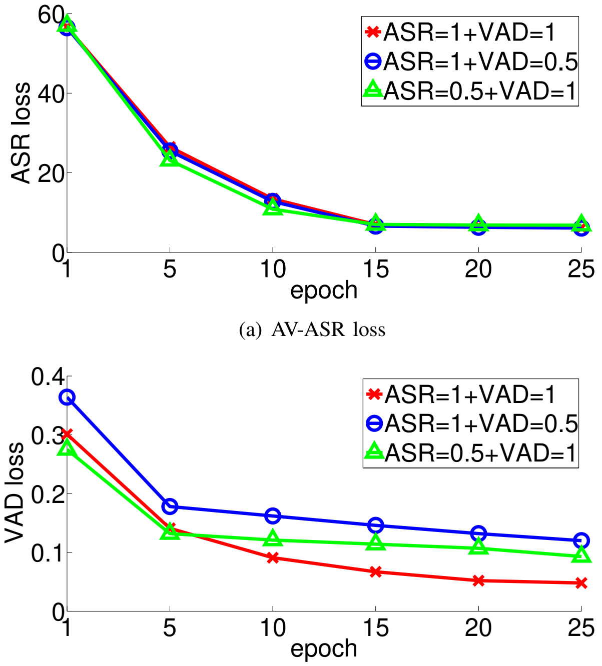
TABLE III
PERFORMANCE OF THE PROPOSED MULTITASK FRAMEWORK WITH TWO DIFFERENT SETTING FOR THE WEIGHTS. WE REPORT THE CER FOR ASR, AND F-SCORE FOR VAD (IDEAL: CLOSE-TAKING MICROPHONE, HD CAMERA; CHALLENGING: MICROPHONE AND CAMERA FROM TABLET). ONE ASTERISK INDICATES THAT ONE SYSTEM IS SIGNIFICANT BETTER THAN THE WORSE SYSTEM FOR THE CORRESPONDING SETTING. TWO ASTERISKS INDICATE THAT THE SYSTEM IS SIGNIFICANT BETTER THAN THE OTHER TWO SYSTEMS (p-VALUE<0.01).

Channel	Env	α_{asr}	α_{vad}	AV-ASR CER [%]	AV-VAD F-score [%]
Ideal	Clean	1.0	1.0	12.5^{**}	95.3^{*}
		1.0	0.5	14.5	94.8
		0.5	1.0	14.5	95.6^{*}
	Noisy	1.0	1.0	15.0^{**}	95.5^{*}
		1.0	0.5	18.9	95.0
		0.5	1.0	17.1^{*}	95.8^{*}
Challenging	Clean	1.0	1.0	15.8^{*}	94.7
		1.0	0.5	16.4	94.4
		0.5	1.0	15.2^{**}	95.1^{*}
	Noisy	1.0	1.0	36.5^{*}	91.9
		1.0	0.5	37.5	91.5
		0.5	1.0	35.6^{**}	91.8

weighting more either the ASR task ($\alpha_{asr} = 1$, $\alpha_{vad} = 0.5$) or the VAD task ($\alpha_{asr} = 0.5$, $\alpha_{vad} = 1$). Table III shows the results of using these weights. As a reference, we also list the results when both weights are equal (i.e., $\alpha_{asr} = \alpha_{vad} = 1$).

Table III shows that the system with equal weights has significantly better performance than the other two settings under the ideal channel. For the challenging channel, the system with the best CER results is the system that weights more the VAD loss. This result can be explained by the stronger regularization imposed by the secondary task (i.e., VAD), improving the system generalization when the testing condition differs from the training condition. Note that the maximum difference in ASR performance in the challenging channel between these two systems is 0.9% (i.e., $\alpha_{asr} = 0.5$, $\alpha_{vad} = 1.0$ versus $\alpha_{asr} = 1.0$, $\alpha_{vad} = 1.0$). Overall, weighing more the ASR loss leads to lower performance for the ASR and VAD tasks. The absolute difference in performance for the AV-ASR task across setting is between 0.6% and 3.9%, which is larger than that absolute difference in performance observed for the AV-VAD task (between 0.3% and 0.8%).

We investigate the reason for the drop in performance when decreasing the value of α_{vad}. Figure 4 shows the loss function during training for the primary and secondary tasks. We observe that the VA-ASR task dominates the loss even when the weights are equally set. The AV-ASR loss is about 190 times larger than the AV-VAD loss at the beginning of the training, and about 120 times larger than the AV-VAD loss at the end of the training. The different is due to the summation in Equation 1, which aggregates the cross-entropy across the entire sequence. By tuning the VAD loss weight from $\alpha_{vad} = 1$ to $\alpha_{vad} = 0.5$, the performance for the secondary task (VAD) is not greatly affected. However, the ASR task is affected, since the regulation from the AV-VAD is weaker, affecting the generalization of the model.



(a) AV-ASR loss

(b) AV-VAD loss

Fig. 4. Loss during training for the primary and secondary tasks. The figure shows that the ASR loss dominates the total loss, which explains the good performance of the proposed approach even when $\alpha_{asr} \leq \alpha_{vad}$.

E. Effect of Facial Landmark Interpolation on Performance

As we discussed in Section III-B, we use linear interpolation to recover frames where IntraFace was not able to estimate the facial landmark. We identify all the videos with at least one frame with interpolated facial landmarks to estimate the effect on ASR performance of using facial landmark interpolation. Since the noise considered in this study mainly affects the acoustic modality, we consider the conditions using clean speech and ideal channel. The performance for the AV-ASR task is a CER of 12.6%. For the AV-VAD task, the F-score is 95.1%. These values are not statistically different from the overall performance reported on Table I (CER of 12.5% for AV-ASR; F-score of 95.3% for AV-VAD). This result indicates that the failures in IntraFace on the videos considered in the analysis do not affect the performance of the systems.

F. Contribution of Modalities

We also developed an audio-based system and a visual-based system to quantify the contributions of each modality, and the benefits of the audiovisual fusion. We did not train the

TABLE IV
EXPERIMENT RESULTS COMPARING UNIMODAL SYSTEMS WITH AUDIOVISUAL SYSTEM. "AV" REFERS TO THE PROPOSED AUDIOVISUAL MTL SYSTEM, "A" REFERS TO THE AUDIO-ONLY MTL SYSTEM, AND "V" REFERS TO THE VISUAL-ONLY MTL SYSTEM. WE REPORT CER FOR ASR, AND F-SCORE FOR VAD. THE TABLE ALSO REPORTS THE NUMBER OF PARAMETERS (IDEAL: CLOSE-TAKING MICROPHONE, HD CAMERA; CHALLENGING: MICROPHONE AND CAMERA FROM TABLET). FONT IN BOLD SHOWS SIGNIFICANT IMPROVEMENT (p-VALUE<0.01).

Channel	Env	Mod	ASR CER [%]	VAD F-score [%]	Param [M]
Ideal	Clean	AV	**12.5**	95.3	2.92
		A	22.0	94.8	2.62
		V	89.0	87.5	1.73
Challenging	Noise	AV	**36.5**	91.9	2.92
		A	40.5	90.7	2.62
		V	89.3	84.3	1.73

unimodal systems from scratch. Instead, we extracted the A-RNN and V-RNN subnets in our model (Fig. 2(b)), using these networks to extract discriminative features for each modality. The parameters of these sub-networks are not modified. On top of these sub-networks, we train a neural network with the same architecture as the AV-RNN subnet (Sec. IV-B) and the MTL layers (Sec. IV-C). For example, the audio-based system is created with the A-RNN, AV-RNN and MTL sub-networks, where only the parameters of the AV-RNN and MTL layers are learned during training using only acoustic features. We only conduct the evaluation on the data from the ideal channel under clean environment, and the data from the challenging channel under noisy environment. These two conditions represent the easiest and hardest conditions in our corpus for our system, respectively.

Table IV lists the performance of the unimodal systems. The results show that the audio-based system has better performance than the visual-based system for these two tasks. Even though the visual-based system performs very poorly for the ASR task (around 89% CER), its information is valuable to complement the knowledge extracted from the audio features. The AV-ASR system has absolute improvements of 9.5% for the clean condition and 4.0% for the noise condition. This result indicates that adding visual information in the proposed audiovisual system increases its robustness, since the visual features are less affected by the acoustic noise [58] (they are still affected by over-articulation due to Lombard effect). For the VAD task, the absolute improvements by adding visual features are limited but consistent for both conditions, with absolute improvements between 0.5% and 1.2%.

We compare the results of the A-ASR and AV-ASR systems to assess the relative improvement per character obtained by adding visual features. Figure 5 shows the results for the challenging channel under noisy conditions. With the exception of "x", all the characters benefit by relying also on facial features. The characters "a", "i" and "u" are the characters that benefit the most with over 20% relative improvements.

The last column of Table IV reports the number of parameters for the audio-only, video-only and audiovisual systems. Adding visual information to an audio-only system only increases the number of parameters from 2.62M to 2.92M (i.e., 0.3M).

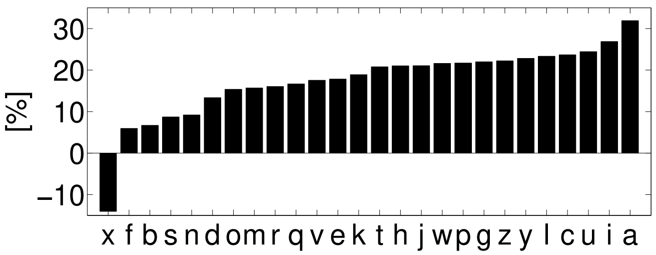


Fig. 5. Relative improvement in ASR performance for each character by considering facial features in addition to audio features (i.e., A-ASR versus AV-ASR systems). The figure reports the results on the test set under challenging condition in noisy environment.

TABLE V
EVALUATION ON THE GRID CORPUS. THE MODELS ARE TRAINED ON THE CRSS-4ENGLISH-14 CORPUS AND TESTED ON THE GRID CORPUS. WE REPORT CER FOR ASR, AND F-SCORE FOR VAD. FONT IN BOLD INDICATES STATISTICALLY SIGNIFICANT IMPROVEMENTS (p-VALUE<0.01).

Model	AV-ASR CER [%]	AV-VAD F-score [%]
MTL	**24.0**	94.3
STL	27.8	-
AE-RNN-MTL	42.7	92.6
AE-RNN-STL	44.1	-

G. Result on the GRID Corpus

We evaluate the performance of the proposed system and the baseline methods on a different corpus to analyze the robustness and generalization of the models under unseen conditions. We consider the GRID corpus [59], which is a publicly available multimodal database. The corpus includes 33 speakers (with audio and video recordings), where each speaker had 1,000 recordings. Each recording consists of a command, color, preposition, letter, digit or adverb. The audio from the corpus was recorded at 50kHz, but we downsampled to 16 kHz for the evaluation. The videos from the corpus were recorded at 25 Hz with a 720 × 576 resolution. The facial landmarks of the videos from the GRID corpus and the template image from the CRSS-4ENGLISH-14 corpus were detected. Then, the faces are normalized by comparing the rigid points on each incoming video frame with the template. For each image, the ROI around the mouth was obtained based on the centroid of all the lips landmarks. The resolution of the ROI is downsampled to 32 × 32. Finally, we use linear interpolation to increase the sampling rate to 29.97 Hz, matching the sampling rate used in the video recording of the CRSS-4ENGLISH-14 corpus. For the audio processing, we perform the same feature extraction process used before using the 16 kHz audio recordings. The evaluation considers models trained on the CRSS-4ENGLISH-14 corpus and tested on the GRID corpus.

Table V lists the results on the GRID corpus. For the ASR task, our proposed approach achieves absolute improvements of 3.8%, 18.7% and 20.1% over the "STL", "AE-RNN-MTL" and "AE-RNN-STL" frameworks, respectively. For the VAD task, our approach is 1.7% better than the "AE-RNN-MTL" framework. The table shows that the proposed approach can outperform the baselines, showing higher performance even when the models are trained with a different corpus. As expected, we observe lower performance than the results reported on the CRSS-4ENGLISH-14 corpus under clean conditions due to the train and test mismatch and the different quality and sampling rate of the videos. In spite of these challenges, the proposed approach achieves a CER of only 24%. The results show that our proposed models are more robust and accurate than the baseline methods when evaluated on a different corpus.

VII. Conclusions

This study proposed a novel end-to-end, multitask learning AV-ASR framework. The approach extended the bimodal recurrent neural network for AV-ASR. The proposed approach combines the feature extraction, modalities fusion, temporal information modeling and classification tasks into one deep neural network. All the parameters of the models are jointly trained toward the desired recognition problem, where the primary task is AV-ASR and the secondary task is AV-VAD. The experimental evaluation demonstrated the benefits of the proposed MTL audiovisual system, leading to consistent reductions in CERs under various channel and environment conditions. Our proposed approach can achieve 12.5% CER under the most ideal scenario and 36.5% under the most challenging scenario. The proposed approach is compared with competitive STL and MTL baselines, obtaining consistent improvements over these systems. Interestingly, the accuracies of the AV-VAD, which is the secondary task, are remarkably high. The system can jointly solve two important problems that are critical in speech-based interfaces: identifying speech segments and recognizing speech content.

The improvements in CER achieved by our end-to-end MTL framework suggest two important observations. First, extracting high level representation directly from the raw inputs can reduce the information loss, which often occurs when using handcrafted features. It can keep the input information as complete as possible as they related to the desired tasks. Second, MTL helps to improve the performance of the system. The VAD task can provide alignment information for the CTC loss, which leads to better models. The MTL models also tend to generalize better across conditions, as the secondary tasks serve as a regularization for the network. The implementation of the approach considers practical settings that can be directly used in real applications. For example, the RNN are implemented with unidirectional LSTM instead of BLSTM to reduce latency in the system. Also, the VAD labels needed to train the system are directly obtained with forced alignment using the transcriptions available to train the ASR task. Therefore, our implementation does not require extra labor to obtain ground truth label for the secondary tasks.

There are several research directions to extend this study. The weights for the primary and secondary tasks are arbitrarily set with the same value. We follow this approach given the computational resources required to train this end-to-end MTL system. We expect better performance by optimizing these hyper-parameters. Furthermore, the results reported in this study rely exclusively on the acoustic models, where the performance can be improved by using a language model. Finally, we are interested in evaluating the proposed audiovisual solutions in actual implementations. We are particularly interested in *human robot interactions* (HRI) situated in noisy acoustic environments.

Acknowledgment

This study was funded by the National Science Foundation (NSF) grant IIS-1718944.

References

[1] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, B. Roomi, and P. Hall, “English conversational telephone speech recognition by humans and machines,” *ArXiv e-prints (arXiv:1703.02136)*, March 2017.

[2] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, “Audio-visual speech recognition,” Workshop 2000 Final Report, Technical Report 764, October 2000.

[3] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior, “Recent advances in the automatic recognition of audiovisual speech,” *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, September 2003.

[4] F. Tao and C. Busso, “Lipreading approach for isolated digits recognition under whisper and neutral speech,” in *Interspeech 2014*, Singapore, September 2014, pp. 1154–1158.

[5] L. Gao, R. Zhang, L. Qi, E. Chen, and L. Guan, “The labeled multiple canonical correlation analysis for information fusion,” *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 375–387, February 2019.

[6] N. El Din Elmadany, Y. He, and L. Guan, “Multimodal learning for human action recognition via bimodal/multimodal hybrid centroid canonical correlation analysis,” *IEEE Transactions on Multimedia*, vol. To Appear, 2019.

[7] L. Pang, S. Zhu, and C. Ngo, “Deep multimodal learning for affective analysis and retrieval,” *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2008–2020, November 2015.

[8] J. Yan, W. Zheng, Q. Xu, G. Lu, H. Li, and B. Wang, “Sparse kernel reduced-rank regression for bimodal emotion recognition from facial expression and speech,” *IEEE Transactions on Multimedia*, vol. 18, no. 7, pp. 1319–1329, July 2016.

[9] N. Li, J. Jain, and C. Busso, “Modeling of driver behavior in real world scenarios using multiple noninvasive sensors,” *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1213–1225, August 2013.

[10] N. Takahashi, M. Gygli, and L. Van Gool, “AENet: Learning deep audio features for video analysis,” *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 513–524, March 2018.

[11] S. Dupont and J. Luettin, “Audio-visual speech modeling for continuous speech recognition,” *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 141–151, September 2000.

[12] P. Wu, H. Liu, X. Li, T. Fan, and X. Zhang, “A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion,” *IEEE Transactions on Multimedia*, vol. 18, no. 3, pp. 326–338, March 2016.

[13] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, J. Chen, J. Chen, Z. Chen, M. Chrzanowski, A. Coates, G. Diamos, K. Ding, N. Du, E. Elsen, J. Engel, W. Fang, L. Fan, C. Fougner, L. Gao, C. Gong, A. Hannun, T. Han, L. Johannes, B. Jiang, C. Ju, B. Jun, P. LeGresley, L. Lin, J. Liu, Y. Liu, W. Li, X. Li, D. Ma, S. Narang, A. Ng, S. Ozair, Y. Peng, R. Prenger, S. Qian, Z. Quan, J. Raiman, V. Rao, S. Satheesh, D. Seetapun, S. Sengupta, K. Srinet, A. Sriram, H. Tang, L. Tang, C. Wang, J. Wang, K. Wang, Y. Wang, Z. Wang, Z. Wang, S. Wu, L. Wei, B. Xiao, W. Xie, Y. Xie, D. Yogatama, B. Yuan, J. Zhan, and Z. Zhu, “Deep speech 2: End-to-end speech recognition in english and mandarin,” in *International Conference on Machine Learning (ICML 2016)*, New York, NY, USA, June 2016, pp. 173–182.

[14] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *International Conference on Machine Learning (ICML 2006)*, Pittsburgh,PA, USA, June 2006, pp. 369–376.

[15] R. Sanabria, F. Metze, and F. De La Torre, “Robust end-to-end deep audiovisual speech recognition,” *ArXiv e-prints (arXiv:1611.06986)*, vol. abs/1611.06986, pp. 1–5, November 2016.

[16] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, July 1997.

[17] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 951–955.

[18] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, December 2018.

[19] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1103–1107.

[20] ——, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 3698–3702.

[21] F. Tao and G. Liu, "Advanced LSTM: a study about better time dependency modeling in emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 2906–2910.

[22] F. Tao, G. Liu, and Q. Zhao, "An ensemble framework of voice-based emotion recognition system for films and TV programs," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 6209–6213.

[23] N. Chen, Y. Qian, and K. Yu, "Multi-task learning for text-dependent speaker verification," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 185–189.

[24] S. Parveen and P. Green, "Multitask learning in connectionist robust asr using recurrent neural networks," in *European Conference on Speech Communication and Technology (EUROSPEECH 2003)*, Geneva, Switzerland, September 2003, pp. 1813–1816.

[25] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, June 2018.

[26] F. Tao and C. Busso, "Gating neural network for large vocabulary audiovisual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1286–1298, July 2018.

[27] X. Fan, C. Busso, and J. Hansen, "Audio-visual isolated digit recognition for whispered speech," in *European Signal Processing Conference (EUSIPCO-2011)*, Barcelona, Spain, August-September 2011, pp. 1500–1503.

[28] F. Tao, J. Hansen, and C. Busso, "An unsupervised visual-only voice activity detection approach using temporal orofacial features," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 2302–2306.

[29] F. Tao, J. L. Hansen, and C. Busso, "Improving boundary estimation in audiovisual speech activity detection using Bayesian information criterion," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 2130–2134.

[30] G. Potamianos, E. Marcheret, Y. Mroueh, V. Goel, A. Koumbaroulis, A. Vartholomaios, and S. Thermos, "Audio and visual modality combination in speech processing applications," in *The Handbook of Multimodal-Multisensor Interfaces, volume 1*, S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, Eds. ACM Books, May 2017, vol. 1, pp. 489–543.

[31] J. Huang and B. Kingsbury, "Audio-visual deep learning for noise robust speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver, BC, Canada, May 2013, pp. 7596–7599.

[32] A. K. Katsaggelos, S. Bahaadini, and R. Molina, "Audiovisual fusion: Challenges and new approaches," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1635–1653, September 2015.

[33] N. Harte and E. Gillen, "TCD-TIMIT: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, May 2015.

[34] S. Chaudhuri, J. Roth, D. Ellis, A. Gallagher, L. Kaver, R. Marvin, C. Pantofaru, N. Reale, L. Reid, K. Wilson, and Z. Xi, "Ava-speech: A densely labeled dataset of speech activity in movies," *arXiv preprint arXiv:1808.00606*, 2018.

[35] D. Dov, R. Talmon, and I. Cohen, "Audio-visual voice activity detection using diffusion maps," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 732–745, April 2015.

[36] S. Thermos and G. Potamianos, "Audio-visual speech activity detection in a two-speaker scenario incorporating depth information from a profile or frontal view," in *2016 IEEE Spoken Language Technology Workshop (SLT)*. San Diego, USA: IEEE, Dec. 2016, pp. 579–584.

[37] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *International conference on machine learning (ICML2011)*, Bellevue, WA, USA, June-July 2011, pp. 689–696.

[38] J. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, Honolulu, HI, USA, July 2017, pp. 3444–3453.

[39] F. Tao and C. Busso, "Aligning audiovisual features for audiovisual speech recognition," in *IEEE International Conference on Multimedia and Expo (ICME 2018)*, San Diego, CA, USA, July 2018, pp. 1–6.

[40] I. Ariav, D. Dov, and I. Cohen, "A deep architecture for audio-visual voice activity detection in the presence of transients," *Signal Processing*, vol. 142, pp. 69–74, January 2018.

[41] R. Ahmad, S. Raza, and H. Malik, "Unsupervised multimodal VAD using sequential hierarchy," in *IEEE Symposium on Computational Intelligence and Data Mining (CIDM 2013)*, Singapore, April 2013, pp. 174–177.

[42] S. Takeuchi, T. Hashiba, S. Tamura, and S. Hayamizu, "Voice activity detection based on fusion of audio and visual information," in *International Conference on Audio-Visual Speech Processing (AVSP 2009)*, Norwich, United Kingdom, September 2009, pp. 151–154.

[43] I. Almajai and B. Milner, "Using audio-visual features for robust voice activity detection in clean and noisy speech," in *European Signal Processing Conference (EUSIPCO 2008)*, Switzerland, Lausanne, August 2008, pp. 1–5.

[44] F. Tao and C. Busso, "Bimodal recurrent neural network for audiovisual voice activity detection," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1938–1942.

[45] S. Petridis and M. Pantic, "Deep complementary bottleneck features for visual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 2304–2308.

[46] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 3652–3656.

[47] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning (ICML 2014)*, Beijing, China, June 2014, pp. 1764–1772.

[48] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, "End-to-end audiovisual speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 6548–6552.

[49] F. Tao and C. Busso, "Audiovisual speech activity detection with advanced long short-term memory," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 1244–1248.

[50] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, "Audio-visual speech recognition with a hybrid CTC/Attention architecture," in *IEEE Workshop on Spoken Language Technology (SLT 2018)*, Athens, Greece, December 2018, pp. 513–520.

[51] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *International Conference on Machine Learning (ICML 2008)*, Helsinki, Finland, July 2008, pp. 160–167.

[52] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013)*, Portland, OR, USA, June 2013, pp. 532–539.

[53] J. Lyons, "Python speech features," https://github.com/jameslyons/python_speech_features, 2016.

[54] F. Tao and C. Busso, "End-to-end audiovisual speech activity detection with bimodal recurrent neural models," *ArXiv e-prints (arXiv:1809.04553)*, pp. 1–11, September 2018.

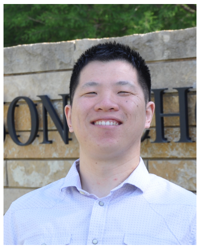
[55] A. Katsamanis, M. P. Black, P. Georgiou, L. Goldstein, and S. Narayanan, "SailAlign: Robust long speech-text alignment," in *Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, Philadelphia, PA, USA, January 2011, pp. 1–4.

[56] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, San Diego, CA, USA, May 2015, pp. 1–13.

[57] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, June 2014.

[58] T. Tran, S. Mariooryad, and C. Busso, "Audiovisual corpus to analyze whisper speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013)*, Vancouver, BC, Canada, May 2013, pp. 8101–8105.

[59] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “An audio-visual corpus for speech perception and automatic speech recognition,” *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, November 2006.



Fei Tao (S’13) received the BS degree in electrical engineering from Beijing Jiaotong University, Beijing (BJTU), China in 2009, the MS degree in Transportation Department from Texas Southern University (TSU), Houston, USA, and the PhD degree at the Electrical Engineering Department of The University of Texas at Dallas (UTD) in 2018. At BJTU, he received the university scholarship from 2005 to 2008. He also received the second prize in the 2008 Beijing College-Student Circuits Design Contest. In 2011, he received the Dwight David Eisenhower President Fellowship for his research in Intelligent Transportation System (ITS) at TSU. His research interests include digital signal processing, speech and video processing, audio visual speech recognition and multimodal fusion.

Carlos Busso (S’02-M’09-SM’13) received the BS and MS degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the PhD degree (2008) in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. He is an associate professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He was selected by the School of Engineering of Chile as the best electrical engineer graduated in 2003 across Chilean universities. At USC, he received a provost doctoral fellowship from 2003 to 2005 and a fellowship in Digital Scholarship from 2007 to 2008. At UTD, he leads the Multimodal Signal Processing (MSP) laboratory [http://msp.utdallas.edu]. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. In 2015, his student received the third prize IEEE ITSS Best Dissertation Award (N. Li). He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). He is the co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research interest is in human-centered multimodal machine intelligence and applications. His current research includes the broad areas of affective computing, multimodal human-machine interfaces, nonverbal behaviors for conversational agents, in-vehicle active safety system, and machine learning methods for multimodal processing. His work has direct implication in many practical domains, including national security, health care, entertainment, transportation systems, and education. He was the general chair of ACII 2017 and ICMI 2021. He is a member of ISCA, AAAC, and a senior member of the IEEE and ACM.