Cross Modal Video Representations for Weakly Supervised Active Speaker Localization

Rahul Sharma, Krishna Somandepalli, and Shrikanth Narayanan, Fellow, IEEE

Abstract—An objective understanding of media depictions, such as inclusive portrayals of how much someone is heard and seen on screen such as in film and television, requires the machines to discern automatically who, when, how, and where someone is talking, and not. Speaker activity can be automatically discerned from the rich multimodal information present in the media content. This is however a challenging problem due to the vast variety and contextual variability in media content, and the lack of labeled data. In this work, we present a cross-modal neural network for learning visual representations, which have implicit information pertaining to the spatial location of a speaker in the visual frames. Avoiding the need for manual annotations for active speakers in visual frames, acquiring of which is very expensive, we present a weakly supervised system for the task of localizing active speakers in movie content. We use the learned cross-modal visual representations, and provide weak supervision from movie subtitles acting as a proxy for voice activity, thus requiring no manual annotations. Furthermore, we propose an audio-assisted post-processing formulation for the task of active speaker detection. We evaluate the performance of the proposed system on three benchmark datasets: i) AVA active speaker dataset, ii) Visual person clustering dataset, and iii) Columbia dataset, and demonstrate the effectiveness of the cross-modal embeddings for localizing active speakers in comparison to fully supervised systems.

Index Terms—Cross-modal learning, weakly supervised learning, multiple instance learning, active speaker localization.

I. INTRODUCTION

TREMENDOUS variety and amounts of multimedia content are created, shared, and consumed everyday, and across the world, with a great influence on our everyday lives. These span various domains, from entertainment and education to commerce and politics, and in various forms; for example, in the entertainment realm these include film, television, streaming, and online media forms. There is an imminent need for creating human-centered media analytics to illuminate the stories being told by using these various content forms to understand their human impact: both societal and economic. Recent efforts to address this need has led to the emergence of computational media intelligence (CMI) [1] which deals with building a holistic understanding of persons, places, and topics involved in telling stories in multimedia, and how they impact the experiences and behavior of individuals and society at large.

Creating such rich media intelligence requires the ability to automatically process and interpret large amounts of media content across modalities (audio, video, language, etc.), each modality with its strengths and limitations to help understand the story being told. The ability to process multiple modalities hence becomes essential to learn robust models for media content analysis. It should be noted that humans concurrently process and experience different aspects of the presented media: sights, sounds, and language use to develop a holistic understanding of the story presented [2]. For example, several studies in psychology and neuroscience have shown evidence for how visual perception in humans is intertwined with other senses such as sound and touch. These mechanisms can be altered even at early stages of development of the primary visual cortex (e.g., [3]). This integration of multiple sensory modalities to holistically perceive visual stimuli is a widely studied field in human psychology, referred to as crossmodal perception [4]. Recently, there have been several works focused on the computationally harnessing the idea of crossmodal perception in the audio-visual domain. Most of these studies use the idea of the naturally existing relations in the audio and the corresponding visual frames, in produced media content [5], [6], [7], [8].

When and where constructs are the fundamental pillars of CMI, for developing a holistic understanding of a scene, which direct to locate the action of interest in time and space. In this paper, we address the problem of visual speech event localization in (Hollywood) movies, which essentially detects speech activity in space (where), signifying active speakers’ faces in the visual frames. Inspired by the cross-modal integration in humans to address the challenges of partial observability and dynamic variability of the audio and visual modalities, we developed a cross-modal neural network that can efficiently fuse the complementary information of the visual and audio modalities to localize a visual speech event effectively.

In our preliminary work [9], we introduced a cross-modal problem formulation for the task of visual voice activity detection. We proposed a 3D convolutional network that observes the raw visual frames of a video segment and predicts the posterior for segment-level audio voice activity detection (VAD). We further established that the learned embeddings were capable of localizing humans in the visual frames. In this work we further
advance the proposed framework for localizing active speakers in space (visual frames). The novel contributions reported include the following:

1) We introduce an enhanced cross-modal architecture consisting of 3D convolutional neural networks (CNNs) and stacked convolutional Bi-LSTMs. This enables the system to capture multi-scale temporal context and introduces an ability to learn hierarchical abstractions in the presented information. The presence of convolutional operations throughout the architecture, in CNNs as well as in Bi-LSTMs, enables the system to preserve the spatiotemporal information, thus making it interpretable at several levels.

2) We present an end-to-end trainable cross-modal system for active speaker localization in visual frames, trained in a weakly supervised fashion. The proposed setup utilizes a multiple instance learning formulation designed for detecting the presence of speech in audio while considering the location of active speaker faces in the visual modality as the key instances.

3) We propose an audio-assisted active speaker detection formulation, which uses the high-level information from the audio stream and integrates them with active speaker posteriors obtained from the visual information, as a post-processing step. Furthermore, we evaluate the system’s performance on three benchmark datasets comprising videos from movies (AVA active speaker dataset [10]), TV shows (Visual person clustering dataset [11]), and a panel discussion (Columbia dataset [12]) and demonstrate performance comparable to fully-supervised methods.

II. RELATED WORK

A. Cross-Modal Learning

There has been a recent surge of studies focused on cross-modal machine perception, especially in media content analysis. The idea of cross-modal learning primarily revolves around modelling one modality guided by another. In [13], the authors target video advertisement classification, using cross-modal autoencoders, reconstructing one modality from the other. In a more recent work by [14], a cross-modal relation-aware network is proposed for audio-visual event localization involving a self-attention mechanism where query is derived from one modality while the key-value pairs the other. Another work [15] targets the problem of fake news detection using a cross-modal residual network, where the text modality guides the attention for learning visual representation and vice-versa. In our earlier work [9], we proposed a cross-modal problem setup for the task of visual VAD involving a hierarchically context-aware network (HiCA) which observes the visual frames and predict the audio VAD labels.

B. Weakly Supervised Object Detection (WSOD)

WSOD refers to the training setup when only image level labels are provided for supervision opposed to bounding box labels in fully-supervised scenarios. Recent research in WSOD can be broadly categorized into two directions, i) Class activation maps (CAMs), and ii) Multiple instance learning (MIL) based setups. CAMs based methods leverage the relationship between CNN embeddings and the class posteriors to compute localization maps. One of the earlier approaches [16] used the idea that the recognition score will drop if the object of interest is artificially masked out in the input image. The idea of CAMs [17] was initially proposed to compute the discriminative image regions for a class of interest in the case of linear prediction layers. Grad CAM [18] was later introduced, generalizing CAMs by using the gradients of the posteriors with respect to the activations of the pertinent layer. Furthermore, GradCAM++ [19], introduced weighted average of pixel-wise gradients to improve the coverage of detections and dealt with multiple occurrences of the same object.

MIL setups pose the input image for classification as a bag of instances where instances are object proposals. In an early attempt [20] a two stream CNN was proposed, one stream to predict bag scores while the other one to compute the instance level scores. Recently [21] proposed a multistage instance classifier (MIDN) to predict the tighter object detection boxes, which is further enhanced to improve coverage of detection by using 2 MIDMs [22]. To alleviate the non-convexity issues associated with MIL [23] proposed to use a combination of smoothed loss functions.

C. Active Speaker Localization

Earlier works [24] in active speaker detection largely focus on using the activity in the lip region available in the visual modality. In another approach [6], [7], [8], [25], [26], authors proposed to use the synchrony between the cropped images of lip regions and the associated audio to determine active speakers. Furthermore, [27] introduced the use of cues from upper body movement to determine an active speaker, which they further refined using personalized voice models [12]. Recently [10] proposed a large scale dataset (AVA active speaker dataset), consisting of movies and the corresponding active speaker annotations along with baseline performance using a supervised framework. Several frameworks have since followed [26], [28], [29], [30], [30], [31] for improving the performance on the AVA dataset. But all these works are restricted to supervised frameworks. To overcome the need of expensive annotations [25], [32] proposed a self-supervised framework trained for the task of audio visual correspondence.

D. Sound Source Localization

The problem of active speaker localization falls within the general domain of sound source localization, but for a particular audio event: speech. The core idea driving the research in this direction is to exploit the existing audiovisual correspondence in the media content. Earlier efforts [33], [34], [35] used canonical correlation analysis to model the audio-visual correspondence. Recent research has been dominated by self-supervised deep learning methods, where researchers try to capture the audio-visual correspondence using various proxy tasks. One such proxy task [8], [36] uses the additive nature of audio and
reconstruct the sound for each pixel by learning a mask for the audio spectrogram. Another proxy task [7] predicts the time alignment of the given audio and video pair. The work by [6] used the audio-visual correspondence to predict a localization score for every pixel and [37] extended the same formulation for object detection. Furthermore [14] proposed a cross-modal attention mechanism for audio event classification and used the learned attention for modeling the localization task. Majority of these works qualitatively established the gained localization ability from the inherent audio-visual correspondence but lacks quantitative evaluation. In this work we present a qualitative as well as thorough quantitative analysis of the acquired localization ability of the visual embeddings.

III. METHODOLOGY

A. Cross-Modal Visual Representations

1) Problem Formulation: The work in this paper is especially motivated by the application of active speaker localization in media content such as entertainment media, notably Hollywood movies. From a computer vision perspective, movie videos are challenging due to the presence of rich variety and high dynamics in the content with potentially multiple variable number of persons in both the foreground and the background. Supervised modeling of such videos requires large amounts of (labeled) data in form of bounding-box annotations. In particular, training an audiovisual system for person localization task in a supervised fashion requires large-scale bounding box annotations, which are tedious and expensive to acquire. Inspired by the recent success of cross-modal representations in understanding media content [14], [15], we formulate our problem in a cross-modal fashion where we model a function of audio modality i.e., talking/non-talking person, by directly observing the visual frames. This helps us in circumventing the widespread issues of drop in performance while jointly modeling multiple modalities against uni-modal systems [38], that arises primarily due to the difference in the rate of generalization for different modalities.

In our preliminary work [9], we trained a cross-modal network for predicting segment-level audio voice activity by using the visual information and established that the learned embeddings implicitly acquired a capability to localize humans in the visual frames. Motivated by the attained localizing ability, in this work we modified the cross-modal formulation described in [9], such that the learned embeddings can localize active speakers in the visual frames. To do so, we propose a modified formulation of the learning task to predict the presence of speech (PoS) for a video segment by observing the visual frames. For a given video segment \( v_i \) of \( t - \text{seconds} \), we define PoS as the step function of the duration of voice activity.

\[
\text{PoS} = \begin{cases} 
1 & \text{duration of voice activity} > 0 \\
0 & \text{duration of voice activity} = 0 
\end{cases}
\]  

(1)

The task of predicting PoS is specifically chosen with a hypothesis that the neural network will assess the active speaker regions in visual frames as the most salient to detect the PoS in the video segment. Segment-wise voice activity labels (VAD) can be ambiguous for video segments with partially present speech. Depending on the definition of segment-wise VAD, the VAD label may be false for video segments with a small fraction of speech, but active speaker faces are still present in the video segment. The PoS labels are more relaxed and help resolve such ambiguity in the video segment-level labels. In our experimental setup, we use data from Hollywood movies for training under this formulation and thus utilize the readily available movie subtitles to acquire the PoS labels involving no manual annotations. The relaxed nature of the PoS labels (compared to VAD) makes it easier to obtain finer labels.

Formally, given a video \( V \), we partition the video into smaller segments \( v_i \) of \( t - \text{seconds} \) each. For each of the \( v_i \), we acquire a label \( y_i \), where \( y_i \) indicating the PoS in the video segment. The network sees \( k \) such small segments at once, and the network is trained for the mapping problem \( v_i \rightarrow y_i \). In the current setup, \( t = 1 \text{ sec} \) and \( k = 10 \).

\[
\{v_i, \ldots, v_{i+k}\} \rightarrow \{y_i, \ldots, y_{i+k}\} \quad y_i \in \{0, 1\} \quad (2)
\]

2) Cross-Modal Network Architecture: To model the visual signal in a cross-modal fashion, in preliminary work [9], we introduced a Hierarchical Context-Aware (HiCA) architecture providing the temporal context at different levels, modeling the short-term context using 3D CNNs and long-term context using BiLSTM. Furthermore, we quantitatively and qualitatively established that the trained representations were selective to human faces and the human body. In this work, we enhance the decentralized temporal context of the HiCA architecture by employing three stacked convolutional Bi-LSTM on top of the 3D CNNs to provide multi-scale temporal context. The introduction of the stacked Bi-LSTMs is motivated by the fact that stacked LSTM networks introduce a hierarchical level of abstractions, as established in various works [39] in the field of Natural Language Processing. The convolutional Bi-LSTMs enable the integrated interpretability for the architecture, since they preserve the spatial and temporal structure of the input. Such a model also enables visualizing the learned representations at different levels of the stacked Bi-LSTMs, allowing to analyze the learned hierarchical abstractions. The elaborated neural network architecture is shown in Fig. 1.

The network is trained on a set of 268 Hollywood movies, released during the period 2014-18. The videos are sampled at 24 frames per second and are lowered in resolution to 180 × 360 pixels. The Presence of speech labels are implicitly obtained using the readily available movie subtitles since they correspond to the human speech dialogues present in the movies. The obtained labels are coarse and do not employ any manual annotations.

The subtitles are first processed to remove the presence of special sounds by removing the content quoted within [ ], / { }. It has been observed that the acquired subtitles are not accurately time aligned with the audio. We used the gentle force aligner, a Kaldi-based tool to align speech and text, which time aligns the

\[1\text{[Online]. Available: https://lowerquality.com/gentle/}\]

\[2\text{[Online]. Available: https://kaldi-asr.org/}\]
The cross-modal architecture with 3D CNNs and stacked convolutional BiLSTM layers. The network observes the raw visual frames and is trained to predict the presence of speech (PoS) activity in audio modality.

Fig. 2. Class activation maps for positive class imposed on the input frames showing the localization ability of the learned embeddings. Sample frames from videos of Row1: AVA, Row2: Friends and Row3: TBBT.

subtitles and audio and provides a confidence score with each alignment. We discard the part of the videos which has not been aligned with high enough confidence (empirically determined). We further compute a binary label for each $t - \text{sec}$ of the video segments using the presence of subtitles as a proxy for presence of speech. To provide a tolerance for subtitle alignment errors, we assign a video segment a positive PoS label only if it has subtitles appearing for more than 10% of the duration.

After pre-processing and time aligning the subtitles with audio, we obtained, on average, nearly 70% of the movie duration with a high enough speech-subtitle alignment confidence score. We used $k = 10\ \text{sec}$ and $t = 1\ \text{sec}$, which were driven heuristically, ensuring that CNNs and LSTMs observe enough temporal context to learn. Our training set consists of nearly 360 hours of video data, which comprises 130 k samples (1.3 million video-label pairs, since each sample consists of 10 pairs). The network has been optimized to minimize the cross-entropy loss using an accelerated SGD optimizer for nearly 1 million iterations for a batch size of 8. The data consisting of the PoS labels for the 268 Hollywood movies will be publicly available to promote research.

3) Visualizing Representations: The utmost factor motivating the use of convolutional networks throughout the cross-modal architecture is the ability of CNNs to enhance the interpretability of the learned embeddings. In this work we use an extension of GRAD CAMs [18] to 3D CNNs to visualize the information learned by the visual embeddings. We first differentiate the output sigmoid score, the posterior $\hat{p}_i$ for PoS, with respect to each of the filters $F^m$ of the pertaining convolutional layer with $m$ filters. The obtained gradients are aggregated across temporal and spatial dimensions to obtain the contribution of each filter towards the presence of speech event as shown in (3) ($Z$ is an averaging factor). The filters of the convolutional layer in consideration are averaged in accordance with the weights computed in (3) ($\alpha_m$), and rectified linearly to obtain the final class activation maps, $C$.

$$\alpha_m = \frac{1}{Z} \sum_i \sum_j \sum_k \frac{\partial \hat{p}}{\partial F^m_{ijk}} \quad C = ReLU \left( \sum_m \alpha_m F^m \right)$$

The proposed cross-modal system consists of 3DCNNs and ConvBiLSTMs, which preserve the temporal and spatial information throughout the network, and thus enable the CAMs to take advantage of this spatio-temporal information to analyze the abstracted information in the various convolutional layers qualitatively. We compute the CAMs for the output from various layers for a given video segment and linearly interpolate them across the temporal and spatial dimensions to obtain the maps matching the spatio-temporal resolution of the given video segment. Fig. 2 shows the positive-class activation maps for selected key-frames from various videos, not seen by the network earlier, and shows that the CAMs provide a non-trivial signal for localizing active speakers. The video segments are present in supplementary materials.

B. Weakly Supervised Active Speaker Localization

In this section, we propose a systematic method to utilize the learned visual embeddings for active speaker localization, in a weakly supervised manner. We present a multiple instance learning (MIL) setup optimized for the proxy task of the presence of speech (PoS) by observing the learned cross-modal embeddings Section III-A2 and face bounding boxes, to model the active speaker faces as the key instances. The setup is inspired by the recent works in weakly supervised object detection [21], [22], [23], [40].
Multiple instance learning (MIL) falls under the domain of supervised learning scenarios, where given labeled bags, each bag consisting of multiple instances, we learn a mapping from bags to labels. Particularly for a binary classification task, the bag is assigned a positive label if at least one of the instances in the bag is positive, and a negative label otherwise. This scenario fits appropriately with our problem formulation described in Section III-A1. We define a small video segment, $v_t$, as a bag, and all the faces appearing during the video segment as instances. We train the system for the presence of speech (PoS) events, thus assigning $v_t$ (bag) a positive label only if at least one of the faces (instances) corresponds to the active speaker.

The proxy task of PoS in the MIL setup is specifically chosen, so as to make it consistent with the earlier cross-modal setup (Section III-A2). With such a setup the cross-modal architecture observes the raw visual frames and is trained for the PoS labels. The output embeddings along with face proposals further become input for the MIL system, which is also trained for the PoS task. This consistency in both the setups’ learning tasks makes them compatible to be trained in an end-to-end fashion. The combined CNN + MIL architecture observes the raw visual frames and predicts the PoS tags. Due to computational constraints, in this work, we restrict to training the two components separately.

1) MIL Problem Formulation: Multiple instance learning architecture while training the MIL system, we introduce a trainable 3D-CNN block which observes the visual embeddings from the HICA architecture and, its output representations further act as an input to the MIL system. This enables fine tuning of the initially learned embeddings for the MIL system. The complete architecture is shown in Fig. 3.

In a recent work [43] it was suggested that in MIL system training, the max-pooling of the instance posteriors, to obtain the bag posteriors, shows a selective behavior highlighting one of the instances among all others. It was also pointed that linear softmax pooling boosts the larger posteriors while suppressing the smaller posteriors at the same time. For the application of active speaker localization, we assume the case of non-overlapping speakers, i.e., there can be at most one active speaker in each frame. This requires the selective behavior of the pooling method, selecting one instance (face), at the frame level. Concurrently, there will likely be more than one frame in the video, consisting of instances of active speakers. Thus we require linear-softmax kind of behavior at inter-frame level pooling, boosting the posterior of the more confident frames.

We propose to use a combination of the two pooling methods, thus pooling the instances in each frame using max-pooling to obtain the frame-level posteriors. We further pool the frame-level posteriors using linear softmax pooling operation to obtain the video level posterior score. Let the instance posterior for the $i$th face in frame $f$ is denoted as $\hat{\rho}_{fi}$. The bag posterior, $\hat{\rho}$, is obtained as shown in (4). We optimize the MIL system for the cross-entropy loss, $\text{Loss}_{MIL}$, between the bag posteriors $\hat{\rho}$ and corresponding PoS labels ($y$).

$$\hat{\rho} = \frac{\sum_f (\max_i \hat{\rho}_{fi})^2}{\sum_f (\max_i \hat{\rho}_{fi})}, \quad \text{Loss}_{MIL} = -y \log(\hat{\rho}) \quad (4)$$

**C. Audio-Assisted Active Speaker Detection**

The proposed MIL framework provides a systematic way to obtain active speaker posteriors for each face bounding box in every frame, derived from just the visual information from the

Fig. 3. The overview of the weakly supervised active speaker localization system. The MIL framework observes the visual representations from cross-modal architecture, and face bounding boxes for each frame and is trained to predict presence of speech in a video segment.
frames. Since active speaker detection is inherently a multimodal task, in this section, we propose a framework to combine the information from the audio modality with the visual modality (the obtained posteriors from the MIL framework) as a post-processing step.

We start with extracting the active voice regions from the audio segment, employing an off-the-shelf voice activity detector [44]. We further partition the obtained voiced regions such that each obtained voice segment consists of a speech from only one speaker and call them speaker-homogeneous speech segments and denote the set of all such speech segments as $S = \{s_n\}$. We use simple heuristics rather than sophisticated neural network systems to obtain speaker-homogeneous speech segments. We partition the voice-active segments by the scene boundaries, inspired by the speaker change being one of the prominent movie-cut attributes [45]. It decreases the likelihood of observing a speaker change in the obtained partitions. We further partition the obtained segments to have a maximum duration of 1 sec. Since the number of speaker changes in a video is constant, partitioning the entire audio into a larger number of segments (smaller duration segments), effectively reduces the fraction of speech segments with a speaker change.

From the visual signal, we extract face tracks using RetinaFace [41] for face detection and SORT [46] for tracking and denote the set of all face tracks as $F_{\text{all}} = \{f_k\}$. Using the start and end time of speech segments, $s_n \in S$, and the obtained face tracks, we collect the set of temporally overlapping face tracks for each $s_n$, and denote them as $F_n = \{f_k\}$. Now we formulate the active speaker detection as a speech-face assignation task, selecting a face track for each speaker-homogeneous speech segment $s_n$ from the set of temporally overlapping face tracks $F_n$, as an active speaker face. We show an example of a speaker-homogeneous speech segment and temporally overlapping face tracks in Fig. 4. We denote the speech-face assignations as:

$$F_n = \{f_k \mid f_k \text{ temporally overlaps } s_n\}$$  \hspace{1cm} (5)

$$\left(s_n \leftrightarrow f_n\right) \in F_n : f_n \text{ active speaker face for } s_n$$  \hspace{1cm} (6)

As a speech-face assignation criterion, we compute an active-speaker likelihood for each face track, $f_k \in F_n$, by computing a score $\alpha_k$ using the average of MIL framework-based posteriors $\rho_i$ for the constituting face bounding boxes. We assign the face track with the maximum score, $\alpha_k$, as the active speaker face for the speech segment $s_n$ if the score is greater than a heuristically driven threshold ($\tau$). On the contrary, the scenario when none of the temporally overlapping face tracks for $s_n$ shows significant active speaker likelihood $\max(\alpha_k) < \tau$, signifies the case with off-screen speakers. The speech-face assignation process is described in Algorithm 1.

$$\alpha_k = \frac{1}{T} \sum_i \rho_i$$  \hspace{1cm} (7)

$$s_n \leftrightarrow f_n \mid f_n = \arg\max_{f_k \in F_n} \alpha_k \text{ and } \alpha_k > \tau$$  \hspace{1cm} (8)

The proposed framework takes advantage of the higher-level information from the audio modality to post-process the obtained face-box level active speaker posteriors from the MIL-based framework. By formulation, speech-face assignation enables two-fold constraints on the active speaker posteriors:

1) It restricts a face to be an active speaker’s face only when speech is present in the audio modality.

2) Speech-face assignation being one-to-one enables the constraint that there can be only one speaker at any time, thus enabling the non-overlapping speech. This constraint is inspired from the nature of entertainment media videos, which are designed to have non-overlapping speech. In Table I we show the fraction of overlapping speech in various datasets and observe a marginal fraction for all.

Additionally, selecting one face-track for the entire speech segment enforces continuity for the active speaker’s face between frames and thus adds a smoothness to the active speaker predictions. We want to emphasize that having only one speaker at any time enables non-overlapping speech constraint: although there can be more than one face present in the frames, only one of them can be an active speaker face.

**Algorithm 1. Speech-Face Assignation Framework.**

1. Obtain $S = \{s_n\}$; // speaker-homogeneous
2. for each $s_n \in S$ do
3.   Obtain $F_n = f_k$; // overlapping
4.   for each $f_k \in F_n$ do
5.     $\alpha_k = \frac{1}{T} \sum_i \rho_i$; // face track scores
6.     end
7.   $f_n = \arg\max_{f_k \in F_n} \alpha_k$;
8.   if $\alpha_n > \tau$ then
9.     $s_n \leftrightarrow f_n$; // assignment
10. end
11. end

| Videos          | AVA active speaker dataset | Friends (VPCD) | TBTT (VPCD) |
|-----------------|---------------------------|---------------|------------|
| Fraction of overlapping speech | 2.31% | 1.39% | 4.42% |

---

*Fig. 4. Example of a speaker-homogeneous speech segment and corresponding temporally overlapping face tracks.*

---
IV. EXPERIMENTS AND EVALUATIONS

A. Qualitative Analysis

Here we evaluate the hypothesis, that the visual embeddings learned to detect the presence of speech in audio modality can localize active speakers in the visual frames, from a qualitative perspective. We visualize the salient regions in the video frames for the positive class i.e., presence of speech event in the audio modality using the methodology described in Section III-A3. Fig. 2 shows the CAMs imposed on the frames in form of heatmaps and demonstrates that the positive class activations situate around human faces with high concentration.

**Multiple faces in a frame:** An approach for generalized sound source localization was recently proposed by [7], where a neural network for audio-visual synchrony was trained, a framework that appears to be closest to our case. This work presents class activation maps for various videos in audio-set [47] and shows that the learned audio-visual representations are selective to human faces and moving lips in case of speech event. But the majority of cases presented in this work consist of just one human face in the frame; moreover, they do not provide extensive quantitative analysis to support the claim. As we specialize the cross-modal embeddings for the presence of speech events, it becomes interesting to see what happens when more than one human face is present in a frame. In Fig. 5, we present the CAMs for selected frames, for videos from various datasets, neither seen by the network during training, corresponding to positive class. The frames are particularly selected to have more than one faces present. The figure illustrate that the learned crossmodal embeddings are able to select the active speaker even when more than one face is present on the frame. The sample videos with imposed CAMs can be found in supplementary material.

**Importance of stacked LSTMs:** In this work, we enhance the HICA [9] architecture with three additional stacked Convolutional-BiLSTM, with an idea that it preserves the spatial and temporal information and achieves hierarchical abstraction. To qualitatively validate the advantages of stacked-ed Convolutional BiLSTM layers, we compare the CAMs at two hierarchical levels, one at the end of 3D convolutional network (Conv_CAMS) and the other at the end of the stacked LSTMs (LSTM_CAMs). In Fig. 6, we present CAMs for the two hierarchical levels under two different scenarios:

1) **Speech event:** We present frames with more than one face present in the frame, from the set of speech events. To make it more informative, we manually marked the active speakers in the frames using a red box. It can be observed that the activations in the case of the Conv_CAMs extend to the non-speaker face as well, while the LSTM_CAMs can correct the activations to concentrate just on the active speaker.

2) **Non-speech event:** In this scenario, we present the CAMs for the frames corresponding to non-speech events. We observed that the Conv_CAMs are concentrating on the faces visible in the frames irrespective of their activity while the LSTM_CAMs can correct the undesired activations, and are selective to speech events.

It can be inferred that the group of 3D convolutional layers is selecting the available faces in the frames and the stacked LSTMs, as they can observe a longer context, are narrowing down to selecting the active speakers.

B. Quantitative Analysis

In the earlier section, we have visually established that the learned cross-modal visual representations can successfully localize the active speakers. This section formally quantifies the performance of the embeddings for active speaker detection using various benchmark datasets. This work targets the active speaker detection problem for videos in entertainment media; thus, we focus on evaluating the performance of the proposed system on datasets consisting of videos from movies and TV shows. We use three widely used datasets for evaluation purposes: i) AVA active speaker dataset [10] (movies), ii) Visual person clustering dataset [11] (TV shows), and iii) Columbia dataset [12] (a panel discussion).

We employ the MIL framework (Section III-B) to obtain face-wise active speaker posteriors, which uses the cross-modal visual representations (Section III-A) obtained by a window-wise inference, with a window length of 10 sec and a stride of 0.5 sec. The proposed audio-assisted system (Section III-C) provides the speech-face assignations utilizing the face-track wise active speaker likelihood computed using the mean of obtained face-wise posteriors of the constituent faces. For the face tracks temporally overlapping with any speech segments, we extend the face track active speaker likelihood score $\alpha_k$ to all the face
Syncnet: Syncnet [25] is closest to this work; it employs an early integration of audio and visual information and models the active speaker detection as a fully-supervised task, training the systems using the A V A active speaker dataset. On the contrary, the proposed system models the visual information without using active speaker annotations and utilizes the higher-level audio information for post-processing. Additionally, the movies in the A V A active speaker dataset are international movies shot earlier; thus, they are different from contemporaneous Hollywood movies (used to train the proposed cross-modal network) in terms of cinematography and illumination conditions, further affecting the system’s performance.

2) Visual Person Clustering Dataset: To evaluate the proposed system’s performance on videos aligning more with entertainment media, we use the videos from Visual Person Clustering Dataset [11], which consists of annotations for widely watched Hollywood TV shows and movies. The annotations constitute the plots for all video datasets and that a significant portion of the videos has more than one face on the screen, thus emphasizing the non-trivial nature of the videos. Particularly for the Columbia dataset, we observe that the majority of the frames has multiple faces, making it a difficult scenario for the trivial baselines.

1) AVA Active Speaker Dataset: The AVA active speaker dataset [10] is one of the few large-scale benchmark datasets for active-speaker detection. It consists of the face bounding-box-wise active speaker annotations for 15 min duration of 161 international movies. We use the official implementation provided by the authors, to compute the mean average precision and report the same for the proposed audio-assisted MIL-based strategy and the mentioned baselines in Table II. We also present the performance of other state-of-the-art methods in Table II. We observe that the proposed audio-assisted MIL-based framework outperforms the random face baseline and the largest face baseline with a significant margin. It even outperforms the solid syncnet baselines; the primary reason is AVA’s noisy audio conditions. In nearly 65% of data, when a speaker is visible, the audio is accompanied either by noise or music [10], which affects Syncnet’s performance. The proposed system relying on visual embeddings is relatively indifferent to the audio noise.

We note that the proposed system is not up to the mark with other state-of-the-art methods, primarily due to the significant difference in the employed strategies. The mentioned systems [10], [29], [30], [48] in Table II employs an early integration of audio and visual information and models the active speaker detection as a fully-supervised task, training the systems using the AVA active speaker dataset. On the contrary, the proposed system models the visual information without using active speaker annotations and utilizes the higher-level audio information for post-processing. Additionally, the movies in the AVA active speaker dataset are international movies shot earlier; thus, they are different from contemporaneous Hollywood movies (used to train the proposed cross-modal network) in terms of cinematography and illumination conditions, further affecting the system’s performance.

To provide insight into the evaluation datasets, we present the distribution of the number of faces in each frame. Fig. 7 shows the plots for all video datasets and that a significant portion of the videos has more than one face on the screen, thus emphasizing the non-trivial nature of the videos. Particularly for the Columbia dataset, we observe that the majority of the frames has multiple faces, making it a difficult scenario for the trivial baselines.

Table II

| Methods          | Strategy      | mAP (%) |
|------------------|---------------|---------|
| Syncnet [25]     | Self supervised | 40.5    |
| Roth et al. [10] | Supervised    | 79.2    |
| Zhang et al. [48]  | Supervised    | 84.0    |
| Alcazar et al. [30] | Supervised   | 87.1    |
| Chong et al. [29]  | Supervised    | 87.8    |
| Alcazar et al. [49] | Weakly supervised | 76.2    |
| Random Face      | Unsupervised  | 47.2    |
| Largest face     | Unsupervised  | 51.0    |
| Audio-assisted Syncnet [25] | Self supervised | 59.7 |
| Audio-assisted MIL | Weakly supervised | 67.3 |

Fig. 7. Distributions of number of faces in each frame for various datasets.
manually verified bounding boxes for instances of primary characters (along with identity) and time stamps of their corresponding speech activity. Although VPCD covers a wide variety of videos from movies (Hidden Figures and About Last Night) and TV shows (Friends, TBBT, Sherlock), the annotations are limited to primary characters. Thus, the videos with more secondary characters have non-exhaustive annotations. Due to the limited number of characters in the TBBT and Friends, the VPCD annotations exhaustively covers all active speaker instances. We thus use the videos from the TBBT (6 episodes) and Friends (25 episodes) for evaluation.

In Table III we report the performance for the baselines and audio-assisted MIL-based framework averaged over the episodes of the TV shows. We note that the proposed system outperforms the random and largest face baselines and the naive Syncnet with a significant margin consistently for all the videos while performing comparably to the audio-assisted Syncnet baseline. We point out that the Syncnet and the proposed system show significantly superior performance for the TV shows over each speaker-homogeneous speech segment offered by the proposed speech-face assignation framework. We observe a significant performance enhancement with the audio post-processing on top of MIL-based face-wise posterior scores from the MIL-based framework as the active speaker scores for each box. Using no further post-processing step, we report the performance in terms of mean average precision.

### C. Ablation Studies

1) **Audio-Assisted Formulations:** In this work, to take advantage of the high-level audio information, we posed the problem of active speaker detection as a speech-face assignment task. To each speaker-homogeneous speech segment, we assign an active speaker face track, from the set of temporally overlapping face tracks, based on the MIL-based posteriors of the constituting face boxes. In this section, we compare the performance of the proposed formulation with two baseline formulations:

- **Visual-only baseline:** Relying on information from just the visual signal, we use the face-wise posterior scores obtained from the MIL-based framework as the active speaker scores for each box. Using no further post-processing step, we report the performance in terms of mean average precision.

- **Audio post-processing:** We obtain the face-wise posterior scores from the MIL-based framework and voice active regions in a video using an off-the-shelf VAD system \cite{44}. As a post-processing step, we make the posteriors of the boxes lying out of VAD regions 0, while keeping the ones in the VAD region unaltered. In addition to the MIL posteriors, this baseline imposes the constraint that active speaker faces lie in the voice active regions only.

In Fig. 8, we present the performance comparison of the above baseline formulations and the proposed speech-face assignment framework. We observe a significant performance enhancement with the audio post-processing on top of MIL-based face-wise posterior scores for all the videos, signifying the importance of the constraint that an active speaker’s face can only be present if there is speech in audio modality. Furthermore, the proposed speech-face assignment framework, consistently for all videos, improves the performance further more significantly for the AVA videos. The observed advantage is due to the additional constraint of having at most one speaker at any time (non-overlapping speech) and the speaker continuity through the speaker-homogeneous speech segment offered by the proposed formulation. We also note the significant improvement in Syncnet performance when assisted with audio using the proposed

### TABLE III

**Performance Comparison of the Audio-Assisted MIL-Based Framework With the Baselines on Videos From VPCD, Reported in % Mean Average Precision (MAP)**

| Methods                        | Friends (25 episodes) | TBBT (6 episodes) |
|--------------------------------|-----------------------|-------------------|
| Random face                    | 52.8                  | 60.8              |
| Largest face                   | 59.6                  | 66.3              |
| Syncnet \cite{25}              | 63.5                  | 70.4              |
| Audio-assisted syncnet \cite{25} | 77.1                  | 80.0              |
| Audio-assisted MIL             | 75.8                  | 81.6              |

### TABLE IV

**Comparison of the Speaker-Wise Weighted F1 Scores (%) for All the Speakers in Columbia Dataset**

| Methods          | Abbas | Bell | Roll | Lieb | Long | Sick | Avg  |
|------------------|-------|------|------|------|------|------|------|
| Chakravarty et al. \cite{12} | 82.9 | 65.8 | 73.6 | 86.9 | 81.8 | 78.2 |      |
| Shahid et al \cite{50} | 89.2 | 88.8 | 85.8 | 81.4 | 86  | 86.2 |      |
| Syncnet \cite{25} | 93.7 | 83.4 | 86.8 | 97.7 | 86.1 | 89.5 |      |
| Aloufas et al. \cite{26} | 92.6 | 82.4 | 88.7 | 94.4 | 95.9 | 90.8 |      |
| S-VNAD \cite{51} | 92.4 | 97.2 | 92.3 | 93.5 | 92.5 | 94   |      |
| Random Face       | 63.8 | 54.2 | 57.1 | 53.3 | 49.5 | 51.9 | 55.0 |
| Largest Face      | 96.9 | 41.7 | 71.4 | 96.6 | 41.3 | 33.8 | 64.0 |
| Audio-assisted MIL| 82.7 | 73.6 | 61.8 | 81.7 | 81.1 | 82.7 | 77.3 |
Fig. 8. Performance comparison of audio-assisted and visual-only formulations for videos from VPCD and AVA, in mean average precision (%).

Fig. 9. Performance comparison of the audio-assisted system (using system VAD), against the oracle speaker-homogeneous speech segments.

system for all the videos (Tables II and III), even though it uses an early-stage audio-visual fusion.

2) Effect of VAD Performance: The proposed audio-assisted framework relies on off-the-shelf VAD system [44] and uses simple heuristics to obtain speaker-homogeneous speech segments. In this section, we explore the impact of the VAD system’s performance on active speaker detection. We compare the performance of the audio-assisted active speaker detection system with an ideal case scenario where we acquire speaker-homogeneous speech segments using the ground truth. Unlike AVA, the annotations in VPCD contain time-stamps for character-wise speech segments, which enables to obtain ideal case speaker-homogeneous speech segments; thus we use videos from VPCD for this study.

In Fig. 9, we show the performance of the proposed audio-assisted system and the one with oracle speaker-homogeneous speech segments. To add further context, we plot the performance of the earlier described visual-only system, utilizing no audio post-processing step. We note that there is a significant difference between the performance of the proposed system and the ideal case scenario, indicating scope for improvement in the active speaker detection performance with the advancements in VAD systems.

Fig. 10. Performance of the audio-assisted framework for 3 groups of face sizes: small (<1%), medium (1–5%) and large (>5%), for VPCD and AVA.

3) Effect of Face Size: Here we investigate the effect of the face sizes in the videos on the system’s performance. We divide all the faces in a video into three sets:

- Small: that occupy less than 1% of the frame.
- Medium: that occupy between 1-5% of the frame.
- Large: that occupy greater than 5% of the frame.

In Fig. 10, we present the proposed audio-assisted system’s performance for face boxes in three groups. We observe that the performance for the small boxes is significantly lower than others for all the datasets. The reason is the lower resolution of the final convolution layer, which is $12 \times 23$. Small boxes constituting lesser than 1% of the frame will end up being represented by at most 2 points in the embedding space. The lack of representative information for small boxes leads to degraded performance. The performance for medium and large boxes is nearly equivalent for the TV shows, while more enhanced for large faces in AVA videos.

V. SUMMARY AND FUTURE WORK

In this paper, we present a cross-modal framework for learning visual representations, capable of localizing an active speaker
in the visual frames. We further formalized a system for active speaker localization, in a weakly supervised manner, requiring no manual annotations. The consistency in the problem formulation for the cross-modal network and the MIL setup makes the system end-to-end trainable. We evaluated the performance of audio-visual speech event localization on the three benchmark datasets comprising a variety of videos, and demonstrated a good active speaker detection performance, provided its weakly-supervised formulation.

The presented system is self-contained in the sense that it can be adapted to any domain in a straightforward manner. To do so, it requires no manual annotations, but just coarse voice activity labels, which can be obtained using the off-the-shelf VAD systems. One of the immediate extensions of our work is to adapt the system for animated content for animated character discovery such as illustrated in Fig. 11. Since the system connects the speech to its spatial source in visual frames, it can be extended to jointly model audio and visual modality for diarization.

REFERENCES

[1] K. Somandepalli et al., “Computational media intelligence: Human-centered machine analysis of media,” Proc. IEEE, vol. 109, no. 5, pp. 891–910, May 2021.

[2] J. Klemen and C. D. Chambers, “Current perspectives and methods in studying neural mechanisms of multisensory interactions,” Neurosci. Biobehavioral Rev., vol. 36 no. 1, pp. 111–133, 2012.

[3] L. Shams and R. Kim, “Crossmodal influences on visual perception,” Phys. Life Rev., vol. 7, no. 3, pp. 269–284, 2010.

[4] K. Schmiedchen, C. Freitag, I. Nitsche, and R. Rübsamen, “Cross-modal interactions and multisensory integration in the perception of audio-visual motion—a free-field study,” Brain Res., vol. 1466, pp. 99–111, 2012.

[5] R. Arandjelovic and A. Zisserman, “Look, listen and learn,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 2843–2851.

[6] P. Tang, X. Wang, X. Bai, and W. Liu, “Multiple instance detection network with online instance classifier refinement,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 618–626.

[7] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2921–2929.

[8] R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 618–626.

[9] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks,” in Proc. IEEE Winter Conf. Appl. Comput. Vis., 2018, pp. 839–847.

[10] J. Y. Feng and A. Vedaldi, “Weakly supervised deep detection networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 2846–2854.

[11] P. Tang, X. Wang, X. Bai, and W. Liu, “Multiple instance detection network with online instance classifier refinement,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2843–2851.

[12] F. Wan et al., “C-MIL: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 9834–9843.

[13] Z. Ren et al., “Instance-aware, context-focused, and memory-efficient object localization with deep networks,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2017, pp. 2843–2851.

[14] P. Tang, X. Wang, X. Bai, and W. Liu, “Multiple instance detection network with online instance classifier refinement,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2843–2851.

[15] J. León-Alcázar, F. C. Heilbron, A. Thabet, and B. Ghanem, “Maas: Multimodal assignment for active speaker detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 265–274.

[16] T. Afours, A. Owens, J. S. Chung, and A. Zisserman, “Self-supervised leaning of audio-visual objects from video,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 208–224.

[17] P. Chakravarty, S. Mirzaei, T. Tuytelaars, and H. Van hamme, “Who’s speaking? audio-supervised classification of active speakers in video,” in Proc. ACM Int. Conf. Multimodal Interact., 2015, pp. 87–90.

[18] Y.-H. Zhang, J. Xiao, S. Yang, and S. Shan, “Multi-task learning for audio-visual active speaker detection,” in Proc. ActivityNet Large-Scale Activity Video Challenge, 2019, pp. 1–4.

[19] M. Everingham, J. Sivic, and A. Zisserman, “Hello! my name is... buffy”–automatic naming of characters in TV video,” Brit. Mach. Vis. Conf., vol. 2, no. 4, p. 6, Sep. 2006.

[20] J. S. Chung and A. Zisserman, “Out of time: Automated lip sync in the wild,” in Proc. Asian Conf. Comput. Vis., 2016, pp. 251–263.

[21] R. Arandjelovic and A. Zisserman, “Look, listen and learn,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 208–224.

[22] J. León-Alcázar, F. C. Heilbron, A. Thabet, and B. Ghanem, “Maas: Multimodal assignment for active speaker detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 265–274.

[23] R. Afours, A. Owens, J. S. Chung, and A. Zisserman, “Self-supervised leaning of audio-visual objects from video,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 208–224.

[24] J. Hershey and J. Movellan, “Audio vision: Using audio-visual synchrony to locate sounds,” in Advances in Neural Information Processing Systems, S. Solla, T. Leen, and K. Müller, Eds., vol. 12, Cambridge, MA, USA: MIT Press, 2000, pp. 814–819. [Online]. Available: https://proceedings.neurips.cc/paper/1999/file/ b618c302e934362ac26c1b280123c22-Paper.pdf

[25] Z. Barzelay and Y. Y. Schechner, “Harmony in motion,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2007, pp. 1–8.

[26] J. S. Chung, “Naver at activitynet challenge 2019–task B active speaker detection (AVA),” 2019, arXiv:1906.10555.

[27] J. León-Alcázar, F. C. Heilbron, A. Thabet, and B. Ghanem, “Maas: Multimodal assignment for active speaker detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 1270–1278.

[28] T. Afours, A. Owens, J. S. Chung, and A. Zisserman, “Self-supervised learning of audio-visual objects from video,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 208–224.

[29] J. S. Chung and A. Zisserman, “Out of time: Automated lip sync in the wild,” in Proc. Asian Conf. Comput. Vis., 2016, pp. 251–263.

[30] R. Afours, A. Owens, J. S. Chung, and A. Zisserman, “Self-supervised learning of audio-visual objects from video,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 208–224.

[31] R. Arandjelovic and A. Zisserman, “Look, listen and learn,” in Proc. Adv. Neural Inf. Process. Syst., 2019, pp. 208–224.

[32] J. S. Chung, “Naver at activitynet challenge 2019–task B active speaker detection (AVA),” 2019, arXiv:1906.10555.
X. Wang, A. Shrivastava, and A. Gupta, “A-fast-RCNN: Hard positive generation via adversary for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 3039–3048.

Y. Wang, J. Li, and F. Metze, “A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling,” in Proc. ICASSP IEEE Int. Conf. Acoust., Speech Signal Process., 2019, pp. 31–35.

H. Bredin et al., “Pyannote.audio: Neural building blocks for speaker diarization,” in Proc. ICASSP IEEE Int. Conf. Acoust., Speech, Signal Process., 2020, pp. 7124–7128.

A. Pardo, F. C. Heilbron, J. L. Alcázar, A. Thabet, and B. Ghanem, “Moviecuts: A new dataset and benchmark for cut type recognition,” in Proc. Eur. Conf. Comput. Vis., 2022, pp. 668–685.

A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, “Simple online and realtime tracking,” in Proc. IEEE Int. Conf. Image Process., 2016, pp. 3464–3468.

J. F. Gemmeke et al., “Audio set: An ontology and human-labeled dataset for audio events,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2017, pp. 776–780.

Y.-H. Zhang, J. Xiao, S. Yang, and S. Shan, “Multi-task learning for audio-visual active speaker detection,” in Proc. ActivityNet Large-Scale Activity Recognition Challenge, pp. 1–4, 2019.

J. L. Alcazar, M. Cordes, C. Zhao, and B. Ghanem, “End-to-end active speaker detection,” in Proc. Conf. Comput. Vis. ECCV, 2022, pp. 126–143, doi: 10.1007/978-3-031-19836-6_8.

M. Shahid, C. Beyan, and V. Murino, “Comparisons of Visual Activity Primitives for Voice Activity Detection,” in Image Anal. Process.: 20th Int. Conf., Trento, Italy, Sep. 9–13, 2019, pp. 48–49, doi: 10.1007/978-3-030-30642-7_5.

M. Shahid, C. Beyan, and V. Murino, “S-VVAD: Visual voice activity detection by motion segmentation,” in Proc. IEEE Winter Conf. Appl. Comput. Vis., 2021, pp. 2331–2340.

Rahul Sharma received the bachelor’s and master’s degrees in electrical engineering from the Indian Institute of Technology Kanpur, Kanpur, India. He is currently working toward the Ph.D. degree in electrical and computer engineering with the Viterbi School of Engineering, University of Southern California, Los Angeles, CA, USA. His research focuses on multi-modal and cross-modal signal processing for long-form media understanding.

Krishna Somandepalli received the master’s degree in electrical and computer engineering from the University of California at Santa Barbara, Santa Barbara, CA, USA, and the Ph.D. degree in electrical and computer engineering from the Viterbi School of Engineering, University of Southern California (USC), Los Angeles, CA, USA. He works on the intersection of affective computing and human-centric understanding in multimedia content with Google Research.

Shrikanth Narayanam (Fellow, IEEE) is currently a University Professor and the Niki & C. L. Max Nikias Chair of engineering with the University of Southern California (USC), Los Angeles, CA, USA, where he is also a Professor of electrical and computer engineering, computer science, linguistics, psychology, neuroscience, pediatrics, and otolaryngology–head & neck surgery, Director of the Ming Hsieh Institute and Research Director of the Information Sciences Institute. Prior to USC, he was with AT&T Bell Labs and AT&T Research. He has authored or coauthored 1000 papers and has been granted 18 U.S. patents. His research and inventions have led to technology commercialization, including through startups he Co-founded: Behavioral Signals Technologies focused on the telecommunication services and AI based conversational assistance industry and Lyssn focused on mental health care delivery, treatment and quality assurance. His research interests include human-centered information processing and communication technologies. He is a Guggenheim Fellow, Member of the European Academy of Sciences and Arts, and Fellow of the National Academy of Inventors, Acoustical Society of America, ISCA, American Association for the Advancement of Science, Association for Psychological Science, Association for the Advancement of Affective Computing and American Institute for Medical and Biological Engineering. He was the recipient of several honors, including the 2015 Engineers Council’s Distinguished Educator Award, Mellon Award for Mentoring Excellence, 2005 and 2009 Best Journal Paper awards from the IEEE Signal Processing Society and was its Distinguished Lecturer during 2010–2011, 2018 ISCA CSL Best Journal Paper Award, and was its an ISCA Distinguished Lecturer during 2015–2016, a Willard R. Zemlin Memorial Lecturer for ASHA in 2017, and Ten Year Technical Impact Award in 2014, and Sustained Accomplishment Award in 2020 from ACM ICMI. He was the inaugural Vice President-Education for the IEEE Signal Processing Society since 2022.