Abstract—Audio and visual signals complement each other in human speech perception, and the same applies to automatic speech recognition. The visual signal is less evident than the acoustic signal, but more robust in a complex acoustic environment, as far as speech perception is concerned. It remains a challenge how we effectively exploit the interaction between audio and visual signals for automatic speech recognition. There have been studies using visual signals as redundant or complementary information to audio input in a synchronous manner. However, human studies suggest another mechanism that visual signal primes the listener in advance, indicating when and which frequency to attend to. To simulate such a visual cueing mechanism, we propose a Predict-and-Update Network (P&U net) for Audio-Visual Speech Recognition (AVSR). In particular, we first predict the character posteriors of the spoken words, i.e. the visual embedding, based on the visual signal. The audio signal is then conditioned on the visual embedding via a novel cross-modal Conformer, that updates the character posteriors. We validate the effectiveness of the visual cueing mechanism through extensive experiments. The proposed P&U net outperforms the state-of-the-art ASR methods on both LRS2-BBC and LRS3-BBC datasets, with the relative Word Error Rate (WER) reductions exceeding 10% and 40% under clean and noisy conditions, respectively.

Index Terms—Predict-and-update, audio-visual speech recognition, early fusion.

I. INTRODUCTION

HUMANS have developed five senses: smell, taste, balance, vision, and hearing as a result of evolution. Among these senses, vision and hearing are primarily involved during social interaction and for effective perception. Human speech perception benefits from combining audio and visual modalities with their unique and complementary characteristics.

It is apparent that multi-modal solutions, such as those combining audio and visual inputs, outperform their unimodal counterparts in speech processing tasks such as speech extraction [1], active speaker detection [2], and emotion recognition [3], [4]. Automatic Speech Recognition (ASR) performance deteriorates in the presence of acoustic noise [5], [6], and so does human speech recognition. For example, the pairs of /m/ and /n/, /b/ and /d/ are acoustically less distinguishable under noise [7]. Visual signals become very useful in a noisy environment because they are not affected by acoustic noise. One of the challenges faced by visual signals is the many-to-one phoneme-to-viseme mapping [8]. In other words, multiple phonemes (up to 13) could be rendered with the same lip movement [8]. The question is how to effectively make use of such an inexact visual cue in speech recognition.

In a human speech perception study [9], Summerfield made three hypotheses about the possible roles of video modality in improving noisy speech intelligibility. 1) Lipreading offers segmental and suprasegmental hints redundant to audio hints. In other words, the phonetic information of speech is equally available in both audio and video modalities. For example, consonants and vowels are segmental, while rhythm, stress, and intonation are suprasegmental. 2) Lipreading offers segmental and suprasegmental hints complementary to audio hints. That is, visual hints provide information unavailable in audio hints, perhaps due to acoustic interference. 3) When one listens, the audio signals (attended speech) and the video signals (lip movements) share common spatial-temporal properties, which may guide listeners to focus on speech signals of interest rather than acoustic interference.

The studies of the first two hypotheses in human speech perception serve as the motivations for implementing of AVSR. As audio and visual signals are either redundant or complementary, they can be fused at the decision level or feature level for improved robustness. At the decision level, we may fuse the decisions made by individual modalities through weighted summation [10], [11] or network [12]. At the feature level, the fusion of either raw features or context representations was studied. A straightforward solution is to first combine raw audio and visual features via concatenation [13], [14], [15] and then apply sequence modeling and decoding. Alternatively, we may first encode the context representations of audio and visual signals...
from the short-time (raw) features, then fuse them, for example, by concatenation with \cite{16, 17} or without \cite{18} audio-visual alignment. Although previous methods motivated by these two hypotheses have achieved great success, they ignore the benefits of the third hypothesis which is suggested by human perception experiments.

In the third hypothesis, a visual cueing mechanism is suggested: lip movement preceding voice \cite{19} can influence hearing at an elementary level, cueing listeners when and on which frequency of speeches one should focus \cite{20, 21}. As well known, closure and release sequentially occur to produce a plosive, where only the latter is audible while both are visible \cite{22}. Thus, visible closure may help predict the coming voice \cite{19}. In a psycho-acoustic experiment, human subjects attempted to distinguish 10 different French syllables pronounced with similar lip movements \cite{23}. Lips were found to move before the arrival of the voice, which prompts listeners to pay attention to sounds, resulting in improved intelligibility. A similar visual cueing effect was reported in the experiment on bird song identification \cite{24}. It was shown that if humans know when a bird song is playing, they can better identify which bird is singing in the presence of acoustic interference without any additional cue, such as a spatial cue. In this case, even simply indicating the onset and offset of the target song in the interference benefits human perception.

Furthermore, in the study of how the video modality helps human speech detection, a high similarity was found \cite{20} between lip-opening area functions and acoustic envelope bands (F2, F3), with up to 165 ms of audio lag, which shows the potential of video to reduce temporal and frequency uncertainty. Further experiments about human speech detection demonstrate that, with video, results of F2-filtered speeches by bandpass are comparable to those of unfiltered speeches, while those of F1-filtered speeches are not comparable \cite{21}. Therefore, preceding the corresponding voice, the video modality cues listeners to focus on when in time and at which frequency on the spectrum.

We are motivated to implement the mechanism of the third hypothesis into a neural architecture. However, it’s unclear which information from the visual modality should serve as the cue and how to utilize this visual cue in the audio modality. Given that the video modality is ambiguous but noise-invariant, while the audio modality is explicit but noise-vulnerable in terms of spoken words, we adopt a predict-and-update network architecture. This approach draws inspiration from the kernel in particle filters \cite{25, 26} and Kalman filters \cite{26}, which are typically used to integrate two sensors with unique properties. The prediction stage uses a sensor measurement that is noise-resistant but less target-related, similar to visual signals in audio-visual speech recognition, to predict the target first. In contrast, another sensor measurement, which is strongly correlated with the target but susceptible to noise, is used in the update stage to refine the target prediction. This mirrors the properties of audio signals. Therefore, we propose using the prediction of spoken characters from the visual signal as the visual cue.

The predict-and-update network, i.e., the P&U net, emulates the visual cueing mechanism in human speech perception by considering the distinct characteristics of the audio and video modalities. As shown in Fig. 2, we first take visual input to predict the probability distributions of spoken text in terms of word characters, which is called visual embedding. We then fuse the visual embedding with the audio features to update the probability distributions. The visual cueing mechanism in the update encoder can be implemented using a factorized-excitation Feed Forward Network (FFN), which does not introduce extra parameters compared to a vanilla FFN. Our contributions are summarized as follows:

1) We propose a P&U net to emulate the visual cueing mechanism in human speech perception while considering the heterogeneous characteristics of audio and visual signals.
2) We design a factorized-excitation Feedforward Network (FFN) to focus on text-related elements within audio features during audio sequence modeling, leveraging visual cues.
3) The proposed audio-visual speech recognition network outperforms other state-of-the-art AVSR methods \cite{18} by a large margin, especially in a noisy environment.
4) We disentangle the third hypothesis of human speech perception and validate that the visual cueing mechanism and early fusion are contributing factors to its advantage.

II. RELATED WORK

We start by reviewing how audio-visual fusion is implemented in the literature to set the stage for this work. The studies of the first two hypotheses in human speech perception have motivated many audio-visual fusion implementations. In particular, audio and visual signals are considered redundant, complementary, and synchronous.

With decision-level fusion, we make the final decisions by combining the decisions of individual modalities \cite{10, 11, 27}, where audio-visual alignment information is not fully exploited. For instance, one can multiply the audio and visual decision probabilities to make a final decision \cite{7}. Note that audio and
video modality have own unique properties, their contributions should be considered separately. In [10], reliability is implemented in making the final decision by weighting between the audio (0.7) and visual (0.3) modality, as audio is more representative in a clean environment. To account for a varying environment, dynamic weights of audio and video modality are applied in [11], [28]. In [28], dynamic weights are controlled by a logistic function of estimated Signal-to-noise ratio (SNR). Recently, decision probabilities are fused with their quality indicators of audio and video modalities (i.e., SNR for audio and facial action unit for video), for example, by a Long Short-Term Memory (LSTM) or fully-connected layer. The decision-level fusion techniques mostly combine text posterior from two modalities for speech recognition. They follow the idea that audio-visual information is redundant or complementary. However, they do not explicitly use synchronization or interaction between audio and video [29]. It is also straightforward to fuse audio and visual signals at the feature level. Fusion can take place between raw features [13], [14], context representations [5], [30], [31], or a mix of both [32], [33]. The fusion of context representations is widely used by concatenation [5], [30] or multiplication [31]. The feature fusion techniques seek to make use of the synchronization information. Some looked into audio-visual alignment before the feature concatenation by varying the hop size of short-time Fourier transform (STFT) [14] or resampling visual signals to the rate of the audio spectrum [13]. Others studied audio-visual interaction by cross-modal attention, where audio context representation queries visual context representation to generate audio-aligned visual representation [16], [34]. This mechanism offers a temporally aligned visual representation to the audio representation. Similarly, visual context representation can query audio context representation [35] to achieve cross-modal interaction. Furthermore, both forms of cross-modal attention are concurrently utilized [17] to maximize audio-visual cross attention. There was also an attempt [36] to map audio and visual raw features to a shared representation to normalize the inherent different modalities. In general, feature-level fusion techniques seek to exploit the audio-visual synchronization property. They have not exploited the asynchronization between audio and visual signals at the time of feature fusion, where the visual signal is processed ahead of the audio signal.

In general, all fusion studies are motivated by the belief that audio and visual signals are redundant and complementary in a synchronous manner. They have not made use of the audio-visual interaction as discovered in the third hypothesis [9] by Summerfield. We are motivated to explore such audio-visual interaction in this work.

III. PREDICT-AND-UPDATE NETWORK

We now propose a novel neural architecture, P&U net, for AVSR, as shown in Fig. 3. In particular, we will elaborate on how the predict-and-update framework emulates the human visual cueing mechanism.

A. Audio and Visual Features

Spectral features are widely used as acoustic features for speech signals in speech recognition [5], [17], and the same applies to AVSR studies [5], [37]. Therefore, we use Short-time Fourier transform (STFT), a representative spectral feature. To compute this feature, we apply a 40-ms filter to the audio signal sequence with a 10-ms hop. Each hop produces a 321-dimensional acoustic feature vector. Then, a Convolutional Neural Network (CNN) module downsamples the sequence of acoustic feature vectors by a factor of four, resulting in 25 frames of acoustic features per second, matching the video sampling rate.

Visual signals are sampled at 25 Frames per Second (FPS). During pre-processing, we crop a 122 × 122 patch from a 224 × 224 visual frame. To balance performance and training effort, we use ResNet [38] to extract visual features [38]. Specifically, we apply a 3D convolutional layer followed by a 2D ResNet-18 [38] to the visual sequences as the visual extractor. The 3D convolutional layer uses a kernel size of 5 × 7 × 7 (time,
height, width). After ResNet-18, we use a global average pooling layer to reduce a 2D visual feature into a 512-dimensional vector.

B. Prediction

We design a predictor to generate a contextual representation, i.e., visual embedding, from a video sequence. The predictor predicts the arrival of sound ahead of time. To this end, an encoder is required to encode the temporal information from a video sequence. There are many options for encoders, such as LSTM [16], [39], Gated Recurrent Unit (GRU) [37], [40], and Transformer [5], [41]. Despite great success, it was argued that these encoders have not explicitly leveraged short-range temporal dependencies [42]. The fully-CNN encoder is one of the solutions for ASR [43] and lip-reading [44]. Recently, Conformer [18], [45] has shown outstanding performance taking advantage of both CNN and the transformer, which take both local and global dependencies into account [18]. We therefore adopt Conformer as our encoders. As shown in Fig. 3(c) and (d), the prediction network consists of some stacked Conformer blocks [45] and a projection head. The Conformer block models temporal dependency in a visual sequence, while the projection head, followed by a softmax function, compresses the visual contextual representation to a low-dimension character posterior. As illustrated in Fig. 1, each Conformer block comprises a FFN layer, a Multi-Head Self Attention (MHSA) module, a convolutional block, and another FFN layer connected sequentially. A layer normalization module is applied before all sub-modules in the Conformer block.

The MHSA module takes as inputs a query feature \( Q \in \mathbb{R}^{T \times d_a} \), a key feature \( K \in \mathbb{R}^{T \times d_k} \) and a value feature \( V \in \mathbb{R}^{T \times d_k} \), where \( T \) is the number of frames in an utterance and \( d_k \) is the dimension of query, key, and value. All three features are injected with position information using relative positional encoding. The MHSA module first maps the set of three features to \( h \) sets according to \( h \) heads and then calculates the scaled dot-product attention for each head, depicted in [41]. The convolutional block consists of two pointwise 1D convolutional layers and a depthwise 1D convolutional layer with a skip connection. It outputs a hidden audio feature \( \tau^c \in \mathbb{R}^{T \times d_a} \), where \( d_a \) is the feature dimension. By combining the convolutional block and the self-attention module [41], the Conformer encoder is capable of modeling both local and global interactions of visual features. The FFN process is formulated as:

\[
FFN(\tau) = W_2 \times \odot (W_1 \tau^T + B_1) + B_2
\]

where \( \tau \in \mathbb{R}^{T \times d_a} \) is the input, \( W_1 \in \mathbb{R}^{d_{ff} \times d_a} \), \( B_1 \in \mathbb{R}^{d_{ff}} \) and \( W_2 \in \mathbb{R}^{d_a \times d_{ff}}, B_2 \in \mathbb{R}^{d_a} \) are the parameters of the first and second linear layers. \( \odot \) is an activation function. In short, FFN firstly expands the feature dimensions of the input \( \tau \) from \( d_a \)-dimensions to \( d_{ff} \)-dimensions and then reduce it back to \( d_a \)-dimensions.

Finally, the projection head maps the output of the encoder into the visual embedding \( \rho \in \mathbb{R}^{T \times C} \), a sequence of \( T \)-frames. Each frame is a \( C \)-dimension feature, representing the posterior probability of target text characters. The prediction module can be pre-trained on a lip-reading task, optimizing \( \rho \) using the CTC and the attention losses, which encourages the predictor outputs feature phonetically relevant to text. The training loss will be discussed in Section III-D.

C. Update

We design an update encoder in Fig. 3(c) to produce an audio-visual context representation using the visual embedding and the raw audio feature. This encoder emulates the visual cueing mechanism. Then the audio-visual context representation is then fed into the decoder and the CTC module.

In the update encoder, the audio-visual fusion is achieved using \( N_{cv} \) cross-modal Conformer blocks. Within these blocks, the same visual embedding cues hierarchical audio features. This design is motivated by that the visual embedding, pre-trained to represent text (character) posterior, has a better understanding of the target text than the hierarchical audio features in the early blocks of the update encoder. This fusion emulates the visual-cueing mechanism. The visual-cued audio features are then processed by \( N_{v} \) vanilla Conformer blocks, which generate audio-visual contextual representation, i.e., the updated text posterior.

We also propose a contrastive architecture by swapping the Conformer and the cross-modal Conformer blocks, as illustrated in Fig. 3(d), to observe the effect between early fusion and late fusion.

The cross-modal Conformer block has the same quantity of parameter as the Conformer [45]. The only difference is that it replaces the original FFN [41] after the convolutional block in the Conformer with a proposed factorized-excitation FFN, as illustrated in Fig. 4. And the factorized-excitation FFN does not include any additional parameters compared to the original FFN, as explained below.
The factorized-excitation FFN is inspired by the factorized layer, which has been implemented in ASR [46] and speech extraction [47]. It is typically used to condition the network on prior external knowledge. Thus, it is a perfect mechanism to take visual cues to strengthen speech encoding. In [46], the factorized layer is used to adapt the ASR model to acoustic conditions with varying speaker genders, identities and noise conditions. So do all $b_k$. Thus, $\mathcal{F}_t$ has the same dimension as the output of the first linear layer of the vanilla FFN,

$$\mathcal{F}_t = \odot(\text{Concat}(SA_{t,1}, SA_{t,2}, \ldots, SA_{t,K}))$$ (5)

where $\odot$ is an activation function.

### D. Loss Function

We optimize the model parameters during training using gradient back-propagation with hybrid loss functions. The model includes mapping modules, the update encoder, the predictor, and the feature extractors. The mapping modules predict a text sequence from the audio-visual context representation.

An AVSR system [5] may employ the CTC loss [52], to force monotonic alignment between the audio-visual context representation and the text sequence, where it is assumed that audio-visual frames are temporally independent of others [53]. The attention-based loss [54] represents another line of thought, which considers inter-frame dependency via an attention mechanism. However, the attention-based loss does not enforce monotonic alignment [30]. To benefit from the best of the two losses, we adopt the hybrid CTC/attention method [18], [55] as the loss function of the P&U net, that delivers promising results [18].

Let us denote $Y \in \mathbb{R}^{L \times C}$ as the ground truth, i.e., the text sequence and $R \in \mathbb{R}^{T \times d_a}$ as the audio-visual context representation, where $L$, $T$, $d_a$ are the length of the text sequence, the number of frames of the audio-visual context representation, and the dimension of the update encoder of the audio-visual context representation, respectively.

In practice, the mapping module of the CTC loss is a linear layer followed by a softmax layer to get the posterior probabilities of $C$ classes for all frames $Y_{ctc} \in \mathbb{R}^{T \times C}$. Afterwards, dynamic programming is used to align the text sequence $Y$ and posterior probabilities $Y_{ctc}$.

The mapping module of the attention loss is a transformer decoder [41], [56] consisting of several transformer blocks. During training, the transformer decoder is adopted to parallel predict all characters in a hypothesis $Y_{att}^t$ of a text sequence from an audio-visual context representation $R$ with $Y$ shifted right. The CE loss is applied to reduce the error between $Y_{att}^t$ and $Y$. During inference, the decoder generates characters one by one to form a character sequence. Unlike the Conformer, the transformer decoder uses absolute positional encoding.

Due to the hybrid training framework, the total loss function is formulated as the weighted sum of CTC loss and attention-based loss, formulated as:

$$L = \lambda p_{ctc}(Y|R) + (1 - \lambda) p_{att}(Y|R)$$ (6)

where $p_{ctc}$ and $p_{att}$ are losses of CTC and attention, $\lambda$ is the coefficient that controls the impact of the two loss components on training.

### E. Predict-and-Update vs. Feature Concatenation

The proposed predict-and-update strategy is different from feature concatenation in terms of the design concept and the actual implementation. First, with predict-and-update, we use.
visual signals to predict the character posterior that is phonetically informing. The fusion takes place between the phonetically informed output distributions and the raw audio features. However, in feature concatenation, the raw features of audio and visual signals are simply put together in the early stage without reference to a text or phonetic content. Second, the visual prediction is achieved through several Conformer blocks that effectively use a video sequence ahead of the arrival of audio sounds, similar to that in the visual cueing mechanism [23], to prime the listener when and on which frequency to focus.

IV. EXPERIMENT

In the visual cueing mechanism of humans speech perception, we observe two properties, 1) a preceding visual signal primes the listener what words are expected, 2) the expected words serve as a cue to the audio signal. The P&U net is designed to study the effect of the above two properties. We first design experiments where the visual embedding primes the update encoder at different stages of audio sequence modelling, i.e., early cueing vs. late cueing. We then compare the proposed methods with some State-of-the-Art (SOTA) systems. As a contrastive system, we also evaluate the traditional feature concatenation technique in comparison with the cueing mechanism. We also perform ablation studies to observe the contributions of the individual modules. All experiments are conducted on large AVSR data sets consisting of uncontrolled spoken sentences.

A. Dataset

LRS2-BBC [5]: The dataset is an audio-visual collection of more than 143K utterances with a 60K vocabulary from BBC programs. All samples are carefully pre-processed by face detection, shot detection, face tracking, facial landmark detection, audio-visual synchronization, forced audio-subtitle alignment, and alignment verification. The facial images are captured from both profile and frontal views. The dataset is split into 4 parts: pre-train, train, val, and test. The pre-train set and the train set are for training, while the val set and the test set are for development and test, respectively. The pre-train set comes with a word-level force alignment. The utterances in the other three parts are shorter than 6 seconds, while a small portion of the utterances in the pre-train set are longer than 6 seconds. The training data is around 224 hours.

LRS3-TED [5]: The dataset contains 164K utterances and have a similar vocabulary size to LRS2-BBC, but with longer utterances. The dataset has a total of 438 hours and is split into 3 parts: pre-train, train, and test. The pre-processing of LRS3-TED is the same as that of LRS2-BBC. Like in LRS2-BBC, the pre-train set also comes with word-level force alignment. However, unlike LRS2-BBC, LRS3-TED has no speaker overlap between the training and test sets, because the data of LRS3-TED is from TED videos and is split by presenters.

We choose character-level tokens as ground truth since words in two datasets are unconstrained. There are 40 output classes, namely \( C = 40 \), including the 26 characters (A-Z), the 10 digits (0–9), the token [space] for separating words, [prime] (‘), [sos] indicates the start or end of an utterance, and [blank] for CTC training.

B. Data Augmentation

In addition, we inject babble noise into waveforms before STFT and implement a method similar to SpecAugment [57] after STFT to augment the data. In detail, we apply two masks each of which is up to 0.4 seconds on the time axis while we randomly mask two frequency bands where each band is less than 1 kHz. The maximum time warp is 5 frames. We generate the babble noise by mixing samples presented in [5]. During training, we follow the SNR distribution of [18], particularly, an uniform distribution over \([\text{No noise}, 20 \text{ dB}, 15 \text{ dB}, 10 \text{ dB}, 5 \text{ dB}, 0 \text{ dB}, -5 \text{ dB}]\).

For video images, patches of \( 112 \times 112 \) are horizontally flipped with 50% random selection.

C. Language Model

Following [18], we adopt a transformer-based [58] character-level Language Model (LM). The training corpus has 16.2 million words, which consists of the transcriptions of LibriSpeech (960 hours) [59], pre-train and train sets of LRS2-BBC [5] and LRS3-TED [5]. Each character is encoded to a 128-D vector without positional encoding. \( d_a \) and \( h \) of the MHSAs and \( d_f \) of the FFN in transformer blocks are 8, 512 and 2048, respectively. Besides, the LM contains 16 transformer blocks. We train the LM by Adam optimizer [60] for 30 epochs with the val set of LRS2-BBC as the development set, learning rate \( 10^{-4} \), and batch size 32. LM training is implemented by ESPnet [61] with a single GeForce GTX 1080 Ti (11GB memory).

D. Pre-Training

As described in Section III-B, we pre-train the predictor via a lipreading task that takes visual frame sequence as input and generates character posteriors. In the predictor, there are 12 Conformer blocks in the visual Conformer encoder, namely \( N_p = 12 \). The hyperparameters are \( d_k = 256, d_v = 256, d_f = 2048, h = 4 \). And the kernel size of the depthwise 1D convolutional layers is 31. Besides, for LRS2-BBC, the visual frontend is trained, which is the same as Fig. 3. But for LRS3-TED, due to limited disk storage and lengthy training procedure if we train the frontend simultaneously, we adopt parameters of the visual frontend from [5] and pre-process image sequences to 512-D visual embeddings via the frozen visual frontend. Because of pre-processing, there is no visual data augmentation for LRS3-TED and we just crop central patches of \( 112 \times 112 \) to the visual frontend in the pre-processing. This setting is consistent in the following AVSR experiments.

In the update encoder, we adopt the factorized-excitation FFN instead of a standard FFN, as shown in Fig. 3(b). As both share the same architecture, we can pre-train an audio Conformer [18] to initialize the update encoder. The audio Conformer serves as an audio-based ASR model but replaces its input filter-bank features or convolutional features by spectrum from STFT. There are 12 Conformer blocks in the ASR Conformer encoder. The hyper-parameters of the ASR encoder are the same as that of the lipreading encoder except for the attention head \( h = 8 \).

The pre-trained lipreading and ASR model share the same decoder architecture as that of the P&U net in Section III-D.
which consists of 6 transformer blocks. The hyper-parameters of MHSA, FFN of the decoder block are set as \(d_k = 256, d_a = 256, d_{ff} = 2048\), \(h = 8\).

We report the performance of the pre-trained lipreading model in Table I. The CTC inference results are solely based on the output of the predictor \(\rho\). The CTC\&LM denotes the system that infers a text sequence by incorporating an external language model of characters. The CTC\&Decoder\&LM further involves the decoder. When the CTC, the decoder, and the LM are all involved during frame-by-frame inference, the output of LM, the decoder, and the CTC module are summed up with their respective weight of \(\psi, \gamma\), and \(1 - \gamma\) to form a final probability distribution. The top-1 character is used as the output. We empirically set \(\gamma = 0.1, \psi = 0.6\) for LRS2-BBC. As the LM is optimized on the val set of LRS2-BBC, there could be a mismatch with LRS3-TED. Therefore, we choose \(\gamma = 0.2, \psi = 0.4\) for LRS3-TED. Moreover, beam search is employed with a width of 20.

It is reported that the CTC inference by the predictor alone achieves a character error rate of 32.2% and 40.1% for top-1 decoding on LRS2-BBC and LRS3-TED datasets. By incorporating the LM and decoder, the lipreading model is further improved. As we use the visual embedding \(\rho\) rather than the top-1 decoding as the visual cue, we believe that the lipreading model provides informative visual cues.

### E. Implementation Details

The predictor and the update encoder are initialized by the parameters of the encoders of the pre-trained lipreading and ASR models, respectively, while the decoder parameters are initialized by the decoder of the lipreading model. For the variant of the visual embedding \(\rho',\) its dimension \(K = 32\) because 32 is the closest number to the number of output classes and meanwhile is the divisor of the hidden vector dimension of vanilla FFN \(d_{ff}\). Consequently, the dimension of audio sub-spaces \(d_a = 64\).

We implement our models on a single GeForce RTX 3090 (24 GB memory). The network is trained by Adam optimizer [60] with \(\beta_1 = 0.9, \beta_2 = 0.98, \text{and } \epsilon = 10^{-9}\). Batch sizes are 8 for LRS2-BBC and 16 for LRS3-TED. Optimizers take a step every 4 and 2 batches for two dataset. The learning rate linearly increases to \(2 \times 10^{-4}\) by 25000 steps and afterwards decreases proportionally to the inverse square root of the step number, which is employed in [41].

For experiments on LRS2-BBC, we mix its pre-train and train sets as the training set and exclude samples with more than 24 seconds. Because the mixed training set contains more than 140K utterances, we process them by virtual epochs, randomly picking 16,384 samples from the mixed set. The network is trained for 500 virtual epochs. And only samples within 6 seconds will be picked at the first 100 epochs, which is similar to curriculum learning [5].

For experiments on LRS3-TED, the virtual epoch, the mixture of pre-train and trainval set, and the curriculum learning are employed as well. The difference is that we only exclude samples of more than 48 seconds, and we train the network for 750 virtual epochs, as LRS3-TED are bigger than LRS2-BBC.

At the inference phase, the attention decoder module, the CTC module and the LM are employed together. We set the same weights as that in Section IV-D, namely, \(\gamma = 0.1, \psi = 0.6\) for LRS2-BBC, \(\gamma = 0.2, \psi = 0.4\) for LRS3-TED.

Besides, we implement the feature concatenation (Feat Concat) as a contrastive model of the P&U net (early4) to show the benefit of cuing mechanism. TheFeat Concat fuses audio and visual features described in Section III-A by simply joining the time-aligned audio and visual features into a single feature. Before the concatenation, two features are normalized to a similar scale. Subsequently, the concatenated audio-visual feature vector is reduced to a 256-D vector by passing through two linear layers, that is then taken by the 12 Conformer blocks (\(d_k = 256, d_a = 256, d_{ff} = 2048, h = 8\)). The loss function of theFeat Concat is the same as the P&U net.

### F. Timing of Audio and Visual Signals

Since speech lags behind lip movement up to 160 ms given different phonemes in human speech [20], we visualize the activation of ASR and lip-reading to show the consistency between human speech and machine speech recognition.

We adopt ‘continued’ as an example in Table II. Plosives consist of closure and release in order, with only the latter being audible while both are visible [22]. Coincidentally, the plosive '/k/', denoted as C in Table II, appears 3 frames (120 ms) earlier in the visual modality than in the audio modality. We can also observe that the lag between the two modalities is variable. N and T in the visual modality lead by 2 frames (80 ms) and 1 frame (40 ms), respectively, compared to the audio modality.

The remaining activation in the visual modality is inaccurate because the visual modality is more ambiguous when transcribing text.

### G. Cueing at Different Stage

As shown in Fig. 3(c) and (d), the order of the cross-modal Conformer blocks and the Conformer block is inter-changeable. There are a total of 12 blocks in the update encoder. In particular, we design an experiment with five configurations: P&U net

| Inference Method | CTC | CTC&LM | CTC&Dec&LM | WER |
|------------------|-----|--------|------------|-----|
| LRS2-BBC         | 32.2% | 24.8%  | 27.2%      | 40.21% |
| LRS3-TED         | 40.1% | 36.5%  | 34.6%      | 51.97% |

Table I

**Activation of ‘continued’ from ASR and Lip-Reading in the Time Domain**

|               | Activation |
|---------------|------------|
| Video         | C, C, O, N, N, T, T, L, N, N, Š, Š, Š, Y, Y, O, U |
| Audio         | Š, Š, Š, Š, C, O, N, T, T, L, N, N, U, U, E, E, D |

\(\beta\) and \(\gamma\) are the blank and the space between two words, respectively.
TABLE III

| Dataset               | Clean | Noisy | Clean | Noisy |
|-----------------------|-------|-------|-------|-------|
| P&U net (early4)      | 3.83% | 9.50% | 3.05% | 8.48% |
| P&U net (middle)      | 4.11% | 11.33%| 3.14% | 9.69% |
| P&U net (late)        | 4.61% | 14.23%| 3.62% | 12.75%|
| P&U net (early8)      | 3.71% | 9.89% | 3.26% | 8.51% |
| P&U net (all)         | 3.87% | 11.62%| 3.52% | 9.78% |

Noisy: under 0 dB babble noise.

TABLE IV

| Transcription         |                  |
|-----------------------|------------------|
| GT                    | THAT’S A PICKLED WALNUT |
| P&U net (early4)      | THAT’S A PICKLED WALNUT |
| P&U net (middle)      | THAT TO PICKLE WALNUT |
| P&U net (late)        | THAT TO PICKLE WALNUT |
| Audio (early4)        | THAT’S A PICKLE WALNUT |
| Video (late)          | THAT’S THE BUILDING |

Audio: pre-trained ar model; video: pre-trained lipreading model; GT: ground truth.

TABLE V

| Transcription         |                  |
|-----------------------|------------------|
| GT                    | INSTEAD OF BEING SOMEONE’S ARM CANDY |
| Early4                | INSTEAD OF BEING SOMEONE’S ARM CANDY |
| Middle                | INSTEAD OF BEING SOMEONE’S FAMILY |
| Late                  | INSTEAD OF BEING SOMEONE’S AUNTIE |
| Audio                 | INSTEAD OF BEING SOMEONE’S ON KENT |
| Video                 | SELF BECAUSE IT WASN’T COMPLICATED |

Audio: pre-trained ar model; video: pre-trained lipreading model; GT: ground truth; p&u net is omitted before (early4), (middle) and (late).

H. Comparative Study

We compare the P&U net and other SOTA systems on LRS2-BBC in Table VI. In [5], two decoding models, namely TM-CTC using CTC and TM-seq2seq using attention are introduced. They are trained under the MVLRS [62] which is a multi-view lipreading dataset but not publicly available, LRS2-BBC, and LRS3-TED. DCM [17] fuses the context representation of two modalities via cross-modality attention, serving as an example of the proposed method with some SOTA methods on the LRS2-BBC dataset.
attention fusion at the context representation level. TDNN [31] is a fully CNN network, different from others utilizing LSTM or transformer. CTC/Attention [30] takes advantage of both CTC and attention decoding models. AV Conformer [18] employed Conformer instead of Transformer, achieving SOTA performance.

We observe that the P&U net (early4) significantly outperforms other systems both under clean and noisy conditions. Particularly, let us use the best-performing AV Conformer (fusion of separate audio and visual context representation) as the baseline, which has the same learnable parameter as ours. The P&U net (early4) achieves a relative WER reduction of 16.3%, i.e., from a 4.3% WER to a 3.6% WER, over the baseline under the clean condition. We also observe a 42.0% relative WER reduction, i.e., from a 15.7% WER to a 9.1% WER, under the noisy condition.

Note that P&U net has slightly fewer parameters than AV Conformer, since a cross-modal Conformer block has the same quantity of parameters as a Conformer block but P&U net utilizes fewer parameters during audio-visual interaction.

To appreciate the contribution of the visual cueing mechanism, we first compare the worst-performing P&U net (late) and the baseline. The P&U net (late) conducts audio-visual fusion after audio features have passed through 8 Conformer blocks and its performance with clean utterance is worse than that of the baseline (4.6% vs. 4.3%). But the P&U net (late) still outperforms the baseline by a 9.6% relative WER (14.2% vs. 15.7%) in the noisy environment, as shown in Table VI. We are convinced that the visual cueing mechanism contributes to the performance gain by P&U net under noisy conditions.

We also compare the P&U net (early4) and theFeat Concat. Both of them adopt an early fusion strategy, but differ in how audio-visual signals are fused, as discussed in Section III-E. As shown in Table VI, the P&U net (early4) outperforms theFeat Concat under both clean (3.6% WER vs. 4.4% WER) and noisy condition (9.1% WER vs. 11.5% WER) by approximately 20% relative WER reduction. We attribute the performance gain to the visual cueing mechanism.

We further compare the systems on LRS3-TED in Table VII. EG-s2s [37] utilizes the visual modality not only in speech recognition but also in speech enhancement. RNN-T [14] serves as an example of the transducer architecture. Attentive Fusion [63] also fuses audio and visual modalities through attention. Unlike DCM [17], which employs two attention blocks (one with audio as the query and another with video as the query), Attentive Fusion designs a network that only utilizes one attention block (with audio as the query). It shows that this approach outperforms the two-attention-block method under high SNR noise conditions.

We observe that the P&U net (early4) outperforms other SOTA methods, especially under the noisy condition. In particular, the P&U net (early4) leads AV Conformer by 10% relative WER (3.0% vs. 3.3%) and 42.4% relative WER (8.3% vs. 14.4%) under the clean and noisy condition, respectively. Besides, the P&U net (early4) also outperformsFeat Concat under both the clean and noisy conditions, which confirms the benefit of cueing mechanism again.

We also visualize the impact of noise at different SNRs in Figs. 5 and 6. It can be observed that the visual modality contributes more to speech recognition under higher SNR noise conditions. Meanwhile, the performance gap between our proposed P&U net (early4) and the state-of-the-art (SOTA) AV Conformer [18] widens as the noise level increases.

*We re-implement the algorithm using inputs, noise, training procedure, LM described in this paper.

---

**TABLE VII**

| Method                  | Training data | Clean   | Noisy   |
|-------------------------|---------------|---------|---------|
| TM-seq2seq[5]           | MVLSR(730)+LR52630.4(632) | 7.2%    | 42.5%   |
| TM-CCT[5]               | MVLSR(730)+LR52630.4(632) | 7.5%    | 27.7%   |
| DCM[17]                 | LR52630.4(538) | 8.8%    | 30.9%   |
| EG-s2d[37]              | LR530.6(474)  | 6.8%    | 25.5%   |
| RNN-T[14]               | YT31k         | 4.5%    | -       |
| Attention Fusion[63]    | LR530.4(438)  | 6.4%    | -       |
| AV Conformer*[18]       | LR530.4(438)  | 3.3%    | 14.4%   |

Audio-only                   LR530.4(438) | 3.3% | 21.3% |
Feat Concat                  LR530.4(438) | 3.3% | 10.1% |
P&P net (late)               LR530.4(438) | 3.6% | 12.8% |
P&P net (early4)            LR530.4(438) | 3.6% | 8.3%  |

Entries in clean and noisy columns indicate WER with external lm. Noisy: original waveform plus 6 dB babble noise.

---

**Fig. 5.** WERs at different SNRs on the LRS2 dataset.

**Fig. 6.** WERs at different SNRs on the LRS3 dataset.
The cosine similarity between the audio-visual context representations $R$ and $R'$ for speech signals at different SNR on the LRS2-BBC dataset.

| Methods                  | 20 dB | 15 dB | 10 dB | 5 dB  | 0 dB  | -5 dB |
|--------------------------|-------|-------|-------|-------|-------|-------|
| P&U net (late)           | 0.994 | 0.987 | 0.973 | 0.949 | 0.919 | 0.923 |
| AV Conformer*[16]        | 0.987 | 0.972 | 0.945 | 0.885 | 0.778 | 0.721 |

Table VIII

![Figure 7](image_url)

Fig. 7. The WER training curves on LRS2-BBC val set in a teacher forcing mode.

### I. Visual Cueing vs. Context Representation Concatenation

To appreciate how the visual cueing mechanism works under noisy condition, we compare the audio-visual context representation $R$, i.e., the output of the update encoder in the P&U net (late), and the fusion of audio and visual context representations in the baseline. We measure the cosine similarity as expressed in (7) between the audio-visual context representations $R$ for speech signals without noise and $R'$ for that with different SNR noise.

$$\theta_{jt} = \frac{R_{jt} \cdot R'_{jt}}{|R_{jt}| \cdot |R'_{jt}|}$$

where $j$ and $t$ indicate an utterance index and a frame index, respectively, $\theta \in [-1, 1]$ denotes the cosine similarity.

We report the average cosine similarities of the two networks across all time frames and utterances on the LRS2-BBC test set in Table VIII. The cosine similarity $\theta$ of both networks approaches 1.0 when SNR $\geq 15$ dB. We also observe that the cosine similarity of the P&U net (late) is consistently higher than that of the baseline. As the SNR decreases, $\theta$ of the baseline deteriorates rapidly, while $\theta$ of the P&U net (late) remains steady. The results in Table VIII account for why the P&U net (late) performs worse than the baseline under the clean condition but otherwise under 0 dB SNR, which shows the benefit of the visual cueing mechanism under the noisy condition.

### J. Initialization by Pre-Trained Models

The P&U net benefits from the pre-trained ASR and lipreading models as described in Section IV-D. Fig. 7 shows the WER as a function of the number of training epochs for the P&U net and the baseline AV Conformer. The P&U nets are initialized by the pre-trained unimodal models. We also compare AV Conformer with random initialization as a contrast. We can see that the AV Conformer converges more quickly with the pre-trained models than without.

### K. Dimension of Visual Embedding

We are interested in the effect of the pre-defined parameters of the factorized-excitation FFN, for example, the number of audio feature subspaces $K$, which also determines the dimension of audio feature sub-spaces $d_l$. To this end, we design an experiment on LRS2-BBC and LRS3-TED with four different $K = 16, 32, 64, 128$, thus, $d_l = 128, 64, 32, 16$ with the P&U net (early4), and report the WER results in Table IX.

It is observed that the dimension $K$ does not significantly affect the results under both clean and noisy conditions and in both LRS2-BBC and LRS3-TED datasets. Therefore, in the rest of the experiments, we have chosen $K = 32$.

### L. Position of the Factorized-Excitation FFN

We design the cross-modal Conformer block by replacing an FFN module with a factorized-excitation FFN. As shown in Fig. 3(a), there are two FFN modules in a Conformer block. We alter the position of the factorized-excitation FFN by 1) replacing the first FFN module, 2) replacing the second FFN module and 3) replacing both FFN modules. All three configurations are implemented on the P&U net (early4) with the audio subspace dimension $d_l = 64$.

As in Table X, we observe that replacing the first FFN performs slightly better than the second one, which confirms the benefit of early cueing. It is observed that replacing both FFN does not achieve performance gain, suggesting that visual cue plays a secondary role in audio-visual speech recognition.

---

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.


**TABLE XI**

|          | Clean | Noisy |
|----------|-------|-------|
| P&U net (early) | 3.63% | 9.13% |
| P&U net (concat) | 4.52% | 10.78% |

M. A Variant of Predict-and-Update Network

To demonstrate the contribution of the factorized-excitation FFN to the proposed P&U net, we design a variant of the P&U net, called P&U net (concat). It concatenates the visual embedding $\rho$ and audio spectral features in Fig. 3 and maps the concatenation to the hidden feature dimension of the Conformer $d_{\text{c}}$. Next, Conformer blocks instead of cross-modal Conformer blocks are utilized in the update step. Since a cross-modal Conformer block has the same parameter quantity as a Conformer block, the P&U net with the factorized-excitation FFN does not have extra parameters compared with the P&U net (concat).

It is observed that the P&U net with the factorized-excitation FFN outperforms the P&U net (concat) regardless of whether noise exists, as shown in Table XI.

V. CONCLUSION

We have proposed a novel end-to-end audio-visual speech recognition network architecture, i.e., the P&U net. This study is motivated by findings in human speech perception, where visual signals prime the listener for expected words before the arrival of audio signals. We hypothesize that the visual cueing mechanism and early fusion are two key factors contributing to effective audio-visual speech recognition. The experiments have validated these hypotheses. The significant advantage of the visual cueing mechanism over simple fusion suggests that the interaction between audio and visual signals is as important as the signals themselves as far as speech recognition is concerned.

REFERENCES

[1] Z. Pan, R. Tao, C. Xu, and H. Li, “Selective listening by synchronizing speech with lips,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 30, pp. 1650–1664, 2022.

[2] R. Tao, Z. Pan, R. K. Das, X. Qian, M. Z. Shou, and H. Li, “Is someone speaking? Exploring long-term temporal features for audio-visual active speaker detection,” in Proc. ACM Int. Conf. Multimedia, 2021, pp. 3927–3935.

[3] Y. Kim, H. Lee, and E. M. Provost, “Deep learning for robust feature generation in audiovisual emotion recognition,” in Proc. IEEE Int. Conf. Audio, Speech Signal Process., 2013, pp. 3687–3691.

[4] Z. Pan, Z. Luo, J. Yang, and H. Li, “Multi-modal attention for speech emotion recognition,” in Proc. Conf. Int. Conf. Speech Commun. Assoc., 2020.

[5] T. Afousar, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, “Deep audio-visual speech recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 12, pp. 8717–8727, Dec. 2022.

[6] I. C. Yadav and G. Pradhan, “Pitch and noise normalized acoustic feature for children’s ASR,” Digit. Signal Process., vol. 109, 2021, Art. no. 102922.

[7] D. W. Massaro and D. G. Stork, “Speech recognition and sensory integration: A 240-year-old theorem helps explain how people and machines can integrate auditory and visual information to understand speech,” Amer. Scientist, vol. 86, no. 3, pp. 236–244, 1998.

[8] L. Cappelletta and N. Harte, “Phoneme-to-Viseme mapping for visual speech recognition,” in Proc. Int. Conf. Pattern Recognit. Appl. Methods, 2012, pp. 322–329.

[9] Q. Summerfield, “Some preliminaries to comprehensive account of audio-visual speech perception,” Hear. Eye: Psychol. Lipreadng, vol. 3, pp. 746–748, 1976.

[10] J. Luettin, G. Potamianos, and C. Neti, “Asynchronous stream modeling for large vocabulary audio-visual speech recognition,” in Proc. IEEE Int. Conf. Audio, Speech Signal Process., 2001, vol. 1, pp. 169–172.

[11] D. Stewart, R. Seymour, A. Pass, and J. Ming, “Robust audio-visual speech recognition under noisy audio-video conditions,” IEEE Trans. Cybern., vol. 44, no. 2, pp. 175–184, Feb. 2014.

[12] W. Yu, S. Zeiler, and D. Kolossa, “Fusing information streams in end-to-end audio-visual speech recognition,” in Proc. IEEE Int. Conf. Audio, Speech Signal Process., 2021, pp. 3430–3434.

[13] D. Serdyuk, O. Braga, and O. Siohan, “Transformer-based video front-ends for audio-visual speech recognition for single and multi-person video,” in Proc. Conf. Int. Speech Commun. Assoc., 2022.

[14] T. Makino et al., “Recurrent neural network transducer for audio-visual speech recognition,” in Proc. IEEE Workshop Autom. Speech Recognit. Understanding, 2019, pp. 905–912.

[15] B. Shi, W.-N. Hsu, K. Lakhota, and A. Mohamed, “Learning audio-visual speech representation by masked multimodal cluster prediction,” in Proc. Int. Conf. Learn. Representations, 2022.

[16] G. Sterpu, C. Saam, and N. Harte, “How to teach DNNs to pay attention to the visual modality in speech recognition,” IEEE/ACM Trans. Audio, Speech Lang. Process., vol. 28, pp. 1052–1064, 2020.

[17] Y.-H. Lee, D.-W. Jang, J.-B. Kim, R.-H. Park, and H.-M. Park, “Audio-visual speech recognition based on dual cross-modality attentions with the transformer model,” Appl. Sci., vol. 10, no. 20, 2020, Art. no. 7263.

[18] P. Ma, S. Petridis, and M. Pantic, “End-to-end audio-visual speech recognition with conformers,” in Proc. IEEE Int. Conf. Audio, Speech Signal Process., 2021, pp. 7613–7617.

[19] E. Z. Golumbic, G. B. Cogan, C. E. Schroeder, and D. Poeppel, “Visual input enhances selective speech envelope tracking in auditory cortex at a ‘cocktail party’,” J. Neurosci., vol. 33, no. 4, pp. 1417–1426, 2013.

[20] K. W. Grant and P.-F. Seitz, “The use of visible speech cues for improving auditory detection of spoken sentences,” J. Acoust. Soc. Amer., vol. 108, no. 3, pp. 1197–1208, 2000.

[21] K. W. Grant, “The effect of speechreading on masked detection thresholds for filtered speech,” J. Acoust. Soc. Amer., vol. 109, no. 5, pp. 2272–2275, 2001.

[22] D. Byrd, “54,000 american stops,” UCLA Work. Papers Phonetics, vol. 83, pp. 97–116, 1993.

[23] J.-L. Schwartz, F. Berthonnier, and C. Savariaux, “Seeing to hear better: Evidence for early audio-visual interactions in speech identification,” Cognition, vol. 93, no. 2, pp. B69–B78, 2004.

[24] L. A. Varghese, E. J. Ozmeral, V. Best, and B. G. Shinn-Cunningham, “How visual cues for when to listen aid selective auditory attention,” J. Assoc. Res. Otalaryngol., vol. 13, no. 3, pp. 359–368, 2012.

[25] A. Doucet, S. Godsill, and C. Andrieu, “On sequential Monte Carlo sampling methods for Bayesian filtering,” Statist. Comput., vol. 10, no. 3, pp. 197–208, 2000.

[26] R. E. Kalman, “A new approach to linear filtering and prediction problems,” J. Basic Eng., vol. 82, pp. 35–45, 1960.

[27] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, “Learning dynamic stream weights for coupled-HMM-based audio-visual speech recognition,” IEEE/ACM Trans. Audio, Speech Lang. Process., vol. 23, no. 5, pp. 863–876, May 2015.

[28] H. Meutzner, N. Ma, R. Nickel, C. Schymura, and D. Kolossa, “Improving audio-visual speech recognition using deep neural networks with dynamic stream reliability estimates,” in Proc. IEEE Int. Conf. Audio, Speech Signal Process., 2017, pp. 5320–5324.

[29] A. K. Katsaggelos, S. Bahadini, and R. Molina, “Audiovisual fusion: Challenges and new approaches,” Proc. IEEE, vol. 103, no. 9, pp. 1653–1655, Sep. 2015.

[30] S. Petridis, T. Stafylakis, P. Ma, G. Tzimiropoulos, and M. Pantic, “Audio-visual speech recognition with a hybrid CTC/attention architecture,” in Proc. IEEE Speech Lang. Technol. Workshop, 2018, pp. 513–520.

[31] Y. Yu et al., “Audio-visual recognition of overlapped speech for the LRS2 dataset,” in Proc. IEEE Int. Conf. Audio, Speech Signal Process., 2020, pp. 6984–6988.

[32] H. Liu, W. Xu, and B. Yang, “Audio-visual speech recognition using a two-step feature fusion strategy,” in Proc. Int. Conf. Pattern Recognit., 2021, pp. 1896–1803.

[33] D. Hu, C. Wang, F. Nie, and X. Li, “Dense multimodal fusion for hierarchically joint representation,” in Proc. IEEE Int. Conf. Audio, Speech Signal Process., 2019, pp. 3941–3945.
Aivasani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, vol. 30, pp. 6000–6010.

X. Zhang, F. Cheng, and S. Wang, “Spatio-temporal fusion based convolutional sequence learning for lip reading,” in Proc. Int. Conf. Comput. Vis. Pattern Recognit., 2019, pp. 713–722.

J. Li et al., “Jasper: An end-to-end convolutional neural acoustic model,” in Proc. Int. Speech Commun. Assoc., 2019, pp. 71–75.

T. Afouras, J. S. Chung, and A. Zisserman, “ASR is all you need: Cross-modal distillation for lip reading,” in Proc. Int. Conf. Comput. Vis. Pattern Recognit., 2020, pp. 14421–14430.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. Int. Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.

K. Cho et al., “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in Proc. 2014 Conf. Empirical Methods Natural Lang. Process., 2014, pp. 1724–1734.

A. Vlasov et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, vol. 30, pp. 6000–6010.

X. Zhang, F. Cheng, and S. Wang, “Spatio-temporal fusion based convolutional sequence learning for lip reading,” in Proc. Int. Conf. Comput. Vis. Pattern Recognit., 2019, pp. 713–722.

J. Li et al., “Jasper: An end-to-end convolutional neural acoustic model,” in Proc. Int. Speech Commun. Assoc., 2019, pp. 71–75.

T. Afouras, J. S. Chung, and A. Zisserman, “ASR is all you need: Cross-modal distillation for lip reading,” in Proc. Int. Conf. Comput. Vis. Pattern Recognit., 2020, pp. 14421–14430.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. Int. Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in Proc. Int. Conf. Mach. Lear., 2011, pp. 689–696.

B. Xu, C. Lu, Y. Guo, and J. Wang, “Discriminative multi-modality speech recognition,” in Proc. 2023 IEEE Int. Conf. Multimedia Expo, 2023, pp. 2627–2632.

J. S. Chung and A. Zisserman, “Lip reading in profile,” in Proc. Brit. Mach. Vis. Conf., 2017, pp. 155.1–155.11.

L. Wei, J. Zhang, J. Hou, and L. Dai, “Attentive fusion enhanced audio-visual encoding for transformer based robust speech recognition,” in Proc. Asia-Pacific Signal Inf. Process. Assoc. Annua. Summit Conf., 2020, pp. 638–643.