Collaborative Viseme Subword and End-to-end Modeling for Word-level Lip Reading

Hang Chen, Qing Wang, Jun Du, Gen-Shun Wan, Shi-Fu Xiong, Bao-Ci Yin, Jia Pan, Chin-Hui Lee

***Abstract*—We propose a viseme subword modeling (VSM) approach to improve the generalizability and interpretability capabilities of deep neural network based lip reading. A comprehensive analysis of preliminary experimental results reveals the complementary nature of the conventional end-to-end (E2E) and proposed VSM frameworks, especially concerning speaker head movements. To increase lip reading accuracy, we propose hybrid viseme subwords and end-to-end modeling (HVSEM), which exploits the strengths of both approaches through multitask learning. As an extension to HVSEM, we also propose collaborative viseme subword and end-to-end modeling (CVSEM), which further explores the synergy between the VSM and E2E frameworks by integrating a state-mapped temporal mask (SMTM) into joint modeling. Experimental evaluations using different model backbones on both the LRW and LRW-1000 datasets confirm the superior performance and generalizability of the proposed frameworks. Specifically, VSM outperforms the baseline E2E framework, while HVSEM outperforms VSM in a hybrid combination of VSM and E2E modeling. Building on HVSEM, CVSEM further achieves impressive accuracies on 90.75% and 58.89%, setting new benchmarks for both datasets.**

***Index Terms*—Lip reading, deep learning, viseme, hidden Markov model, multitask learning**

I. Introduction

Lip reading aims to recognize spoken content based solely on visual information derived from the speaker's lip movements. This emerging and challenging field lies at the intersection of computer vision and natural language processing and plays a key role in various applications in different domains. For example, in public security, lip reading is crucial for detecting facial forgery and determining liveness [1]. In human-computer interactions, lip reading lays the foundation for silent speech control systems [2], expanding the frontiers of interactive information. In medicine, lip reading significantly benefits people with hearing impairments and those with vocal cord injuries [3]. Moreover, advances in lip reading have also driven audio-visual speech recognition development [4], [5], [6] and enhancement [7], providing an effective method to complement video input in challenging acoustic environments.

Based on the recognition target, lip reading can be categorized into two subtypes: word-level and sentence-level. This study primarily focuses on word-level lip reading (hereafter referred to as "lip reading"). Traditional lip reading systems [8] typically extract handcrafted features from lip-centered video and then adopt a statistical model to decode the features into word-level text. Recently, deep learning has catalyzed remarkable advancements in lip reading. Contemporary state-of-the-art (SOTA) systems [9], [10] typically employ trainable DNNs to classify lip videos into predefined word categories, leveraging extensive training datasets [11], [12]. Subsequent research has focused primarily on improving model backbones [13], [14], [15]. Nevertheless, [16], [14] demonstrated that this end-to-end (E2E) framework is not robust to frame-level noise. Additional boundary annotations can alleviate this issue but have substantial costs. Furthermore, addressing noisy frames from pauses and stutters, among other impediments, has proven challenging with word boundaries alone.

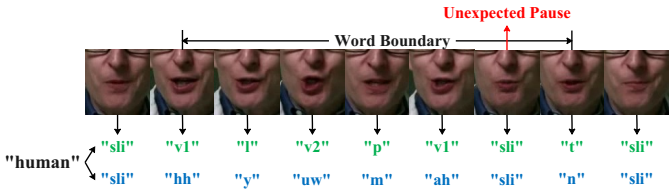


Fig. 1. Illustration of the subword modeling framework. The word label is transformed into a sequence of phone subwords (depicted in blue) or viseme subwords (depicted in green). The system precisely assesses frame-specific contributions by linking the lip frame to its corresponding subword. Notably, compared with phones, visemes offer enhanced distinctiveness.

In response to this challenge, we integrated a subword modeling framework [17] to assist in discerning noisy frames. As illustrated in Fig. 1, the core concept begins by utilizing pronunciation knowledge to convert the word label into a sequence of subwords, followed by modeling an association between each lip frame and its corresponding subword. Diverging from prior research, which focused on refining model backbones, subword modeling gauges the influence of every lip frame on the final decision, thus enhancing lip reading interpretability and generalizability. Although phones are the most common subwords, it is essential to acknowledge that not all phones can be reliably distinguished based on lip movement alone. This is illustrated by the similarity in lip shapes between "hh" and "ah", as shown in Fig. 1. Thus, we advocate adopting visemes [18] as the basic unit. Visemes, which are typically formed by clustering several phones with visually indistinguishable appearances, are more suitable for lip reading and are the core of the proposed viseme subword modeling (VSM) framework.

Furthermore, we comprehensively compare the prediction

H. Chen, Q. Wang and J. Du are with the National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei, Anhui, P. R. China (e-mail: ch199703@mail.ustc.edu.cn; qingwang2@ustc.edu.cn; jundu@ustc.edu.cn). G.-S. Wan, S.-F. Xiong, B.-C. Yin, J. Pan are with the iFlytek Research, Hefei, Anhui, P.R.China (e-mail: gswan@iflytek.com; sfxiong@iflytek.com; bcyin@iflytek.com; jiapan@iflytek.com). C.-H. Lee is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA (e-mail: chl@ece.gatech.edu).

outcomes between the VSM and E2E frameworks. This investigation reveals the complementary aspects of the E2E and VSM frameworks, particularly concerning speaker head movements. Building on this insight, we introduce a pioneering hybrid framework termed hybrid viseme subword and end-to-end modeling (HVSEM). This innovative design seamlessly integrates both VSM and E2E branches into a unified model, facilitating joint training through multitask learning (MTL). However, the parallelization structure in HVSEM fails to capture the commonalities and distinctions among frameworks, which can be mutually beneficial. Accordingly, we further propose a novel collaborative viseme subword and end-to-end modeling (CVSEM) framework to exploit the hierarchical relationship between words and subwords. Specifically, a state-mapped temporal mask (SMTM) module is inserted between the E2E and VSM branches. The SMTM takes frame-level posterior probabilities from the VSM branch as inputs and outputs a Boolean mask to the E2E branch. This mechanism filters unimportant or noisy hidden representations, compelling the E2E branch to make decisions based on more crucial time steps. Notably, all the proposed frameworks, including VSM, HVSEM, and CVSEM, are model-agnostic and applicable to arbitrary lip reading baseline models. Experiments confirm the superior performance and generalizability of the proposed frameworks across various model backbones. The main contributions of our study can be outlined as follows:

1) A novel viseme subword modeling (VSM) framework is proposed that decomposes a word into a sequence of viseme subwords and models each lip frame in association with its corresponding viseme subword.
2) The complementary nature of the E2E and VSM frameworks concerning speaker head movements is identified, and a novel hybrid viseme subword and end-to-end modeling (HVSEM) framework is proposed that exploits this complementarity through multitask learning (MTL).
3) A novel collaborative subword and end-to-end modeling (CVSEM) framework is proposed that enhances the synergy between frame-level and segment-level classification tasks by incorporating an SMTM module.
4) The effectiveness and generalizability of VSM, HVSEM, and CVSEM are confirmed through a series of experiments involving various model backbones and SOTA results are achieved on the widely recognized LRW and LRW-1000 datasets.

The source codes are publicly available[1].

The remainder of the paper is organized as follows. Section II introduces related works. Section III and Section IV describe our proposed frameworks, including VSM, HVSEM and CVSEM. Section V analyses the experimental results. Finally, we summarize our findings in Section VI.

II. Related Work

A. Lip Reading

Conventional lip reading systems typically begin by extracting hand-engineered features from a lip-centered video and then decoding the features into word-level text. Feature extractors commonly fall into two main types: one directly extracts concise features from the pixel values of the lip region through dimension reduction, such as discrete cosine transform (DCT) [19] and maximum likelihood linear transformation (MLLT) [20]. The other type adopts parameters comprising an utterance model based on the contour of the lips, including articulatory features (AFs) [21] and the active contour model (ACM) [22]. Decoders frequently employ hidden Markov models (HMMs) [23] and support vector machines (SVMs) [24]. Further details regarding these traditional methods are discussed in [8].

Recently, deep learning-based lip reading systems [11], [9], in which a DNN replaces separate feature extractors and decoders, have attracted increasing attention. [10] proposed a model backbone consisting of a modified version of ResNet34 followed by a 2-layer bidirectional long-short term memory (BiLSTM) network and partitioning of the whole model into two relatively independent modules: (1) a visual front end that extracts compact and practical visual embeddings from a sequence of lip frames and (2) a sequence back end that aggregates temporal context information to improve recognition performance. Subsequent research has focused mainly on architectural designs of the visual front end and sequence back end.

In terms of designing a visual front end, early approaches typically integrated a shallow 3D convolutional neural network (CNN) with a deeper 2D CNN such as ResNet34 [25], [12], aligned with the frontend in [10]. Several modifications that build upon this foundation have been introduced to improve the architecture, including the spatiotemporal fusion module [13], squeeze-and-extract module [14], time shift module [26], hierarchical pyramidal convolution [15], innovative alternating spatiotemporal and spatial convolution [27] and EfficientNet [28]. In addition, [29] introduced a pseudo3D residual convolution-based front end, which demonstrated the effectiveness of the deep 3D CNN in visual embedding extraction. Moreover, [30] took advantage of deep 3D and 2D CNNs via a parallel two-stream structure with a 2D ResNet-34 and a 52-layer DenseNet-3D [12]. Additional features, such as optical flow [31] and learnable deformation flow [32], have been introduced to expand the two-stream model to further enhance lip reading. Beyond traditional CNNs, graph convolution networks (GCNs), including the raw lip key point graph [33] and adaptive semantic-spatio-temporal GCN (ASST-GCN) [34], have emerged as promising avenues for lip reading research.

In the context of sequence back end design, recurrent neural networks (RNNs) have emerged as foundational components in several approaches. BiLSTM [25], [12] and bidirectional gated recurrent units (BiGRUs) [35], [36] have been widely employed to capture both global and local temporal information. Notably, [37] introduced a fully convolutional architecture consisting of a multiscale temporal convolutional network (MS-TCN) that offered the advantages of faster convergence speed and longer temporal memory. Subsequent works on this topic have made further progress. [38] proposed a densely connected temporal convolutional network (DenseTCN) to

[1]https://github.com/coalboss/CVSEM

TABLE I
THE MAPPING BETWEEN VISEMES AND PHONES.

Description	Viseme	Phone
Lip-rounding based vowels	v1	aa, ah, ao, aw, er, hh, oy
	v2	ow, uh, uw
	v3	ae, ay, eh, ey
	v4	ih, iy
Alveolar-semivowels	l	l, r, y
Alveolar-fricatives	z	s, z
Alveolar	t	d, n, t
Palato-alveolar	sh	ch, jh, sh, zh
Bilabial	p	b, m, p
Dental	th	dh, th
Labio-dental	f	f, v
Velar	g	g, k, ng, w
Silence	sil	sil

capture even more robust temporal features. Additionally, [39] introduced a depthwise separable temporal convolutional network (DS-TCN) to reduce computational costs. Other widely used back end modules include transformers [40] and conformers [41]. Nevertheless, compared to RNN-based and TCN-based back ends, these approaches are more prone to overfitting for small training sets.

However, in these E2E systems, a common approach employs a global average pooling (GAP) layer followed by a linear classifier to derive word posterior probabilities from the output of the sequence back end. The GAP layer assumes that each frame contributes equally to the final decision, making the system susceptible to interference from frame-level noise. As [16], [14] showed, including irrelevant information, such as video frames beyond the actual word boundaries, can significantly degrade a model's performance. One intuitive solution is to introduce additional annotations to mark word boundaries, although this can be a costly endeavor. Consequently, many researchers have proposed automatic methods to address this challenge, including global and local mutual information maximization constraints [42], self-attention-based consensus modules [15] and variational temporal masks [43]. Furthermore, [44] introduced phones as modeling units and proposed a synchronous bidirectional learning (SBL) framework. However, not all phones can be distinguished based on lip movement. Accordingly, [45], [46] proposed a visual-to-audio memory network. This network learns audio representations during training and uses visual input to retrieve stored representations, complementing any information gaps during inference.

B. Viseme

Phonology describes a phone as "the smallest distinct linguistic unit capable of altering meaning" [47]. For speech comprehension, it is necessary to differentiate sounds produced by a speaker as a sequence of separate phones. However, not all phones are distinguishable by observing the movements of visible articulators, such as the mouth, face, or parts of the teeth and tongue. For example, consider the lip representations for "b" and "p", which appear identical to the naked eye, with the only distinguishing factor being the vibration of the vocal folds during phoneme articulation.

A viseme [48] is commonly defined as a speech unit that looks the same during pronunciation. A group of visemes usually consists of numerous phonemes that share an indistinguishable visual appearance. Many mappings from phonemes to visemes have been proposed and investigated [49], [18], [50]. Table I presents a commonly used viseme set [18] and its correlated mapping to phones. Among them, seven consonant visemes were identified, namely, the bilabial, labiodental, dental, palato-alveolar, palatal, velar, and two alveolar visemes. For example, "p", "b" and "m" constitute the well-known bilabial viseme. Lip rounding during vowel formation defined the remaining four vowel visemes and an alveolar-semivowel one, whereas one viseme was devoted to the silence, i.e., "sil". Using visual examples, we can dissect text into visually more precise and well-structured subwords. This simplifies the mapping between basic units and lip frames, making modeling more straightforward and conducive. This approach helps circumvent overfitting issues arising from the ambiguity associated with phones.

C. Multitask Learning

Physiology research [51] suggests that the macaque brain uses the same early processing layers for different tasks rather than segregating and addressing tasks in isolation. These findings have inspired researchers to investigate multitask learning (MTL) [52], a paradigm that seeks to enhance generalization by exploiting domain-specific information in the training signals of all desired tasks. In the deep learning era, MTL involves crafting networks that can acquire shared representations from diverse supervisory signals. This approach not only reduces memory requirements and speeds up inference due to inherent layer sharing, but also more crucially [53], enhances performance by leveraging tasks that provide complementary information or serve as mutual regularizers [54]. [55] provided a survey encompassing feature learning, low-rank task clustering, task relation learning, and decomposition approaches for MTL. Additionally, [56] benchmarked various optimization techniques across three MTL datasets. Notably, the MTL has not been extensively explored in word-level lip reading tasks.

III. VISEME SUBWORD MODELING

Our proposed VSM framework, in both context-independent (CI) and context-dependent (CD) settings, is detailed below.

A. Context-independent VSM

Given a sequence of lip frames denoted as $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_T]$ covering T frames, each $\boldsymbol{x}_t \in \mathcal{R}^{H \times W \times 1}$ represents a grayscale lip image. Lip reading can be formulated as a Bayesian decision problem:

$$\hat{z} = \arg\max_{z \in \mathcal{Z}} p(z|\boldsymbol{X}) = \arg\max_{z \in \mathcal{Z}} p(\boldsymbol{X}|z)p(z) \quad (1)$$

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3390148

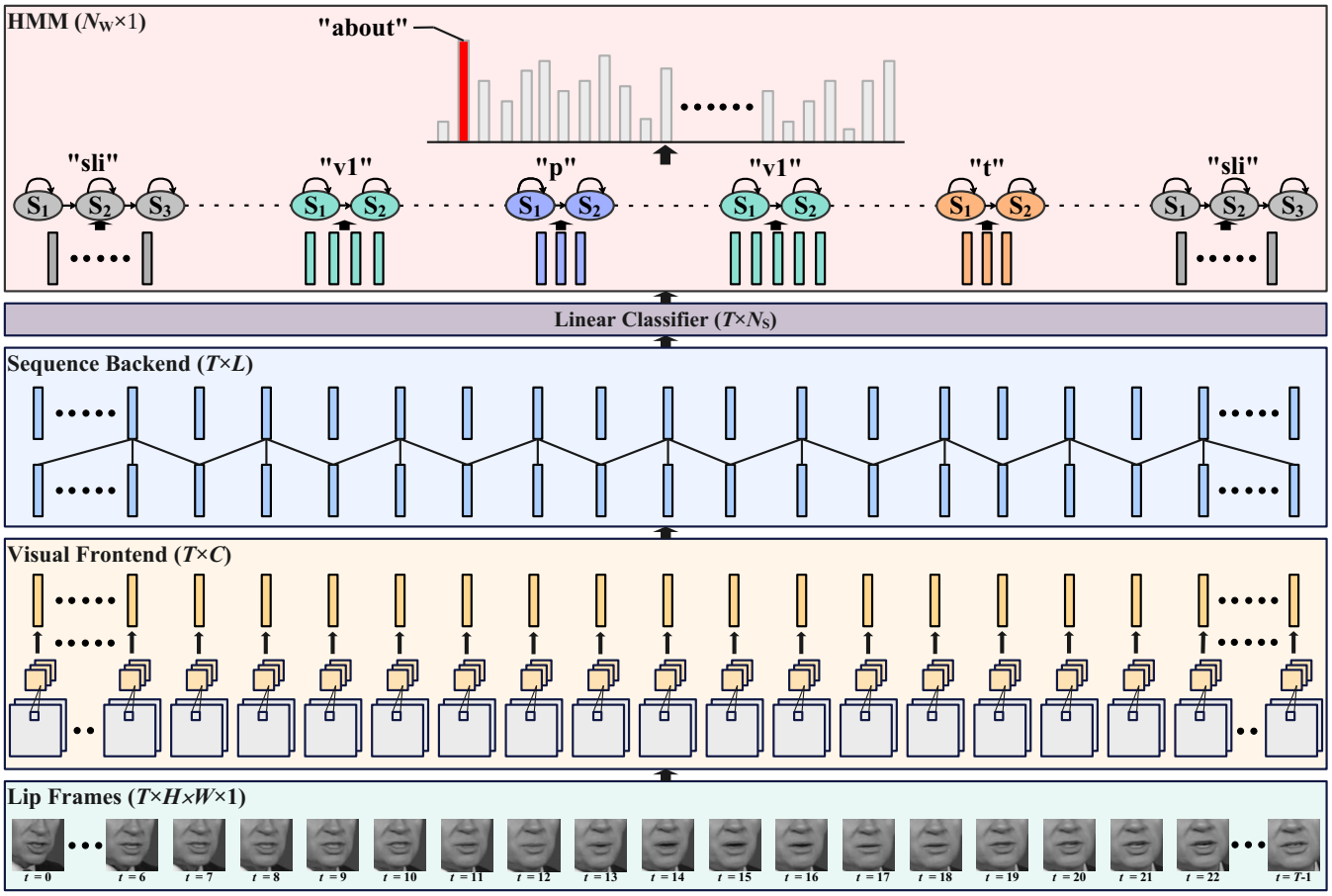


Fig. 2. Illustration of the proposed VSM. A word is decomposed into a sequence of viseme subwords. Each viseme subword is represented by a left-to-right HMM with several hidden states. Each lip frame is assigned to an underlying state corresponding to the word HMM.

where $p(z|\boldsymbol{X})$ is the posterior probability of z given $\boldsymbol{X}$, $p(z)$ is the prior probability of z, which is estimated from the training set, and $p(\boldsymbol{X}|z)$ is the conditional probability of X given z. $\mathcal{Z}$ represents the set of all candidate N_{W} words, and H and W denote the length and width of the lip frame, respectively. Here, we use $H = W = 96$ by default.

As illustrated in Fig. 2, CI-VSM initiates with pronunciation knowledge to convert the word label into a sequence of viseme subwords. Each viseme subword is modeled by employing a left-to-right HMM with multiple states. With HMMs, $p(\boldsymbol{X}|z)$ can be decomposed at the frame level, considering the corresponding hidden state sequence $\boldsymbol{s} = [s_1, \cdots, s_t, \cdots, s_T]$:

$$p(\boldsymbol{X}|z) = \sum_{\boldsymbol{s}\in\mathcal{S}} \left[\pi(s_1) \prod_{t=2}^{T} a_{s_{t-1},s_t} \prod_{t=1}^{T} p(\boldsymbol{x}_t|s_t) \right] \tag{2}$$

$$= \sum_{\boldsymbol{s}\in\mathcal{S}} \left[\pi(s_1) \prod_{t=2}^{T} a_{s_{t-1},s_t} \prod_{t=1}^{T} \frac{p(s_t|\boldsymbol{x}_t)p(\boldsymbol{x}_t)}{p(s_t)} \right] \tag{3}$$

where $\mathcal{S}$ is the hypothesis space of all the candidates. $\pi(s_0)$ is the initial state probability, a_{s_{t-1},s_t} is the state transition probability from frame $t-1$ to t, $p(s_t)$ is the prior probability of state s_t estimated from the training set and $p(\boldsymbol{x}_t)$ is the prior probability of the lip frame $\boldsymbol{x}_t$, which is set to 1. $p(s_t|\boldsymbol{x}_t)$ represents the posterior probability of $\boldsymbol{x}_t$ given s_t, which is estimated by a trainable DNN.

As in existing model backbones, the visual front end $\mathcal{F}_V$ initially takes $\boldsymbol{X}$ as input to derive a vector $\boldsymbol{e}_t \in \mathbb{R}^C$ for each time step, referred to as visual embedding $\boldsymbol{E} = [\boldsymbol{e}_1, \cdots, \boldsymbol{e}_t, \cdots, \boldsymbol{e}_T]$:

$$\boldsymbol{E} = \mathcal{F}_{\mathrm{V}}(\boldsymbol{X}) \tag{4}$$

where C signifies the number of visual embedding channels. We opt for a default value of $C = 512$ in this context. Next, the sequence back end $\mathcal{F}_{\mathrm{S}}$ models the temporal dynamics and outputs the hidden representation $\boldsymbol{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_t, \cdots, \boldsymbol{y}_T]$:

$$\boldsymbol{Y} = \mathcal{F}_{\mathrm{S}}(\boldsymbol{E}) \tag{5}$$

where $\boldsymbol{y}_t \in \mathbb{R}^L$ represents the hidden representation for each time step. L denotes the channel number and is set as the default value of 512 here. Finally, $\boldsymbol{Y}$ is sent to the linear classifier $\mathcal{F}_{\mathrm{L}}^{\mathrm{state}}$ for final prediction $p(s_t|\boldsymbol{x}_t)$:

$$p(s_t|\boldsymbol{x}_t) = \mathcal{F}_{\mathrm{L}}^{\mathrm{state}}(\boldsymbol{y}_t) = \mathrm{SoftMax}(\boldsymbol{y}_t\boldsymbol{W}^{\mathrm{state}} + \boldsymbol{b}^{\mathrm{state}}) \tag{6}$$

where $\mathcal{F}_{\mathrm{L}}^{\mathrm{state}}$ encompasses fully connected and SoftMax layers, with $\boldsymbol{W}^{\mathrm{state}} \in \mathbb{R}^{L\times N_{\mathrm{S}}}$ and $\boldsymbol{b}^{\mathrm{state}} \in \mathbb{R}^{N_{\mathrm{S}}}$ being the weights and the bias of the fully connected layer, respectively. Here, N_{S} represents the total number of HMM states, given by:

$$N_{\mathrm{S}} = 12 \times N_{\mathrm{S}}^{\mathrm{nonsil}} + N_{\mathrm{S}}^{\mathrm{sil}} \tag{7}$$

where $N_{\mathrm{S}}^{\mathrm{nonsil}}$ refers to the number of HMM states corresponding to the nonsilent subword segments, while $N_{\mathrm{S}}^{\mathrm{sil}}$ indicates the number of HMM states corresponding to silence.

CI-VSM training involves minimizing cross-entropy:

$$\mathcal{L} = -\sum_{j=0}^{N_{\mathrm{B}}-1}\sum_{t=0}^{T^j-1}\log p(s_t^j|\boldsymbol{x}_t^j) \tag{8}$$

where j represents the sample index, T^j is the frame number of the j-th sample and N_B signifies the minibatch size employed in training.

To obtain the frame-level HMM state labels required by Eq. (8), we first trained an E2E model that conforms to the HPConv+MS-TCN architecture described in [15]. This model was trained using the Adam optimizer, starting with a learning rate 0.0003 and a weight decay set to $10e-4$. We used a standard cosine scheduler to regulate the learning rate effectively during the 80-epoch training period. This scheduling strategy gradually reduced the learning rate to a minimum of $10e-7$, ensuring a thorough and progressive training process. In addition, to minimize the likelihood of overfitting, a consistent dropout rate of 0.2 was applied to all model configurations.

Next, the pretrained E2E model compresses a lip frame $\boldsymbol{x}_t$ into a lip embedding $\hat{\boldsymbol{y}}$, which is a 256-dimensional vector, using its visual front end. We construct a conventional Gaussian mixture model (GMM)-HMM system using these lip embeddings. Within this framework, the probability output of $\boldsymbol{x}_t$ given the state s_t, denoted as $p(\boldsymbol{x}_t|s_t)$ in Eq. (2), is computed as follows:

$$p(\boldsymbol{x}_t|s_t) = \sum_{m=1}^{M}\left\{\frac{\omega_m}{\sqrt{(2\pi)^L|\boldsymbol{\Sigma}_m|}} \times \exp\left[-\frac{1}{2}(\hat{\boldsymbol{y}}_t-\boldsymbol{\mu}_m)^\top\boldsymbol{\Sigma}_m^{-1}(\hat{\boldsymbol{y}}_t-\boldsymbol{\mu}_m)\right]\right\} \tag{9}$$

where ω_m, $\boldsymbol{\mu}_m$, and $\boldsymbol{\Sigma}_m$ denote the weights, mean vector, and covariance matrix, respectively, for the m-th Gaussian component within the output distribution of state s_t, which is composed of M Gaussian components. Typically, $\boldsymbol{\Sigma}_m$ is a diagonal covariance matrix. The parameters of the GMM-HMM system are optimized via maximum likelihood estimation. For state prior estimation, HMM transition probabilities and GMM weight, mean, and covariance parameters, we use iterative techniques such as the Baum-Welch or Viterbi training algorithms, as described in [57].

Ultimately, the trained GMM-HMM aligns the frame-level lip embeddings with the corresponding HMM states. This process, known as force alignment, resembles a decoding operation, yet it operates within the specific confines of all visemes associated with the word-level label. A thorough and detailed exposition of the force alignment is comprehensively presented in Sec. III-B of [58]. Moreover, [59] offers an accessible tool, providing a practical means for researchers to implement this technique directly in related studies.

B. Context-dependent VSM with State Tying

Lip movements are influenced by physical muscle movements, such as momentum, resulting in modifications of the current lip shape either for the following viseme or based on the previous viseme, referred to as "synchronous articulation". As a response, we utilize CD-viseme subwords to model this behavior explicitly. The testing and training process for the CD-VSM mirrors that of the CI-VSM, as described in the previous section. However, the notable increase in the total number of states poses a challenge regarding trainability. Given a context length of N_{C}, the calculation for the total number of states N_{S} is as follows:

$$N_{\mathrm{S}} = 13^{2N_{\mathrm{C}}}\times(12N_{\mathrm{S}}^{\mathrm{nonsil}}+N_{\mathrm{S}}^{\mathrm{sil}}) \tag{10}$$

Even in the simplest scenario for $N_{\mathrm{C}}=1$, the total number of HMM states N_{S} surpasses 4500 with $N_{\mathrm{S}}^{\mathrm{nonsil}}=2$ and $N_{\mathrm{S}}^{\mathrm{sil}}=3$. Similar HMM states were tied to strike a balance between trainability and accuracy.

State tying is achieved using a binary decision tree, where an algorithm automatically generates questions for each tree node based on the data. In this process, all the states are initially placed in the root node of the tree. Subsequently, by applying a selected question from the generated question set, the node is divided into two child nodes to maximize the increase in log-likelihood. This splitting operation is then repeated to construct the binary tree in a top-down manner until a predefined threshold is reached. Finally, a merge operation of leaf nodes uses a bottom-up approach, utilizing a minimum priority queue. This merging operation is guided by assessing the decrease in log-likelihood to achieve the desired number of tied states.

Given one nonleaf node with the set of HMM states $\mathcal{S}_{\mathrm{f}}$, $\gamma(\mathcal{S}_{\mathrm{f}})$ is the log-likelihood of S generating the N_{O}-frame observation sequence of feature vectors $[\hat{\boldsymbol{y}}_0,\cdots,\hat{\boldsymbol{y}}_{N_{\mathrm{O}}-1}]$. Relying on the assumptions that all tied states adhere to a single Gaussian distribution characterized by mean $\boldsymbol{\mu}(\mathcal{S}_{\mathrm{f}})$ and variance $\boldsymbol{\Sigma}(\mathcal{S}_{\mathrm{f}})$ and that tying states do not alter the frame/state alignment, $\gamma(\mathcal{S}_{\mathrm{f}})$ can be approximated as follows:

$$\gamma(\mathcal{S}_{\mathrm{f}}) = -\frac{1}{2}\sum_{n=0}^{N_{\mathrm{O}}-1}\sum_{s\in\mathcal{S}_{\mathrm{f}}} q(\hat{\boldsymbol{y}}_n|s)[L\ln(2\pi)+\ln|\boldsymbol{\Sigma}(\mathcal{S}_{\mathrm{f}})| + (\hat{\boldsymbol{y}}_n-\boldsymbol{\mu}(\mathcal{S}_{\mathrm{f}}))^\top\boldsymbol{\Sigma}(\mathcal{S}_{\mathrm{f}})^{-1}(\hat{\boldsymbol{y}}_n-\boldsymbol{\mu}(\mathcal{S}_{\mathrm{f}}))] \tag{11}$$

where $q(\hat{\boldsymbol{y}}_n|s)$ represents the observation probability of $\hat{\boldsymbol{y}}_n$ at the n-th frame, generated by the state s. The estimates for $\boldsymbol{\mu}(\mathcal{S}_{\mathrm{f}})$ and $\boldsymbol{\Sigma}(\mathcal{S}_{\mathrm{f}})$ can be obtained as follows:

$$\phi(\mathcal{S}_{\mathrm{f}}) = \sum_{n=0}^{N_{\mathrm{O}}-1}\sum_{s\in\mathcal{S}_{\mathrm{f}}} q(\hat{\boldsymbol{y}}_n|s) \tag{12}$$

$$\boldsymbol{\mu}(\mathcal{S}_{\mathrm{f}}) = \sum_{n=0}^{N_{\mathrm{O}}-1}\sum_{s\in\mathcal{S}_{\mathrm{f}}}\frac{q(\hat{\boldsymbol{y}}_n|s)\hat{\boldsymbol{y}}_n}{\phi(\mathcal{S}_{\mathrm{f}})} \tag{13}$$

$$\boldsymbol{\Sigma}(\mathcal{S}_{\mathrm{f}}) = \sum_{n=0}^{N_{\mathrm{O}}-1}\sum_{s\in\mathcal{S}_{\mathrm{f}}}\frac{q(\hat{\boldsymbol{y}}_n|s)(\hat{\boldsymbol{y}}_n-\boldsymbol{\mu}(\mathcal{S}_{\mathrm{f}}))(\hat{\boldsymbol{y}}_n-\boldsymbol{\mu}(\mathcal{S}_{\mathrm{f}}))^\top}{\phi(\mathcal{S}_{\mathrm{f}})} \tag{14}$$

Using Eq. (14), we have the following derivative for the last

term in Eq. (11):

$$\begin{aligned}
&\sum_{n=0}^{N_{\mathrm{O}}-1}\sum_{s\in\mathcal{S}_{\mathrm{f}}} q(\hat{\boldsymbol{y}}_n|s)(\hat{\boldsymbol{y}}_n-\boldsymbol{\mu}(\mathcal{S}_{\mathrm{f}}))^{\top}\boldsymbol{\Sigma}(\mathcal{S}_{\mathrm{f}})^{-1}(\hat{\boldsymbol{y}}_n-\boldsymbol{\mu}(\mathcal{S}_{\mathrm{f}}))\\
&=\sum_{n=0}^{N_{\mathrm{O}}-1}\sum_{s\in\mathcal{S}_{\mathrm{f}}} q(\hat{\boldsymbol{y}}_n|s)\mathrm{Tr}\{(\hat{\boldsymbol{y}}_n-\boldsymbol{\mu}(\mathcal{S}_{\mathrm{f}}))^{\top}\boldsymbol{\Sigma}(\mathcal{S}_{\mathrm{f}})^{-1}(\hat{\boldsymbol{y}}_n-\boldsymbol{\mu}(\mathcal{S}_{\mathrm{f}}))\}\\
&=\sum_{n=0}^{N_{\mathrm{O}}-1}\sum_{s\in\mathcal{S}_{\mathrm{f}}} q(\hat{\boldsymbol{y}}_n|s)\mathrm{Tr}\{\boldsymbol{\Sigma}(\mathcal{S}_{\mathrm{f}})^{-1}(\hat{\boldsymbol{y}}_n-\boldsymbol{\mu}(\mathcal{S}_{\mathrm{f}}))(\hat{\boldsymbol{y}}_n-\boldsymbol{\mu}(\mathcal{S}_{\mathrm{f}}))^{\top}\}\\
&=\mathrm{Tr}\left\{\boldsymbol{\Sigma}(\mathcal{S}_{\mathrm{f}})^{-1}\sum_{n=0}^{N_{\mathrm{O}}-1}\sum_{s\in\mathcal{S}_{\mathrm{f}}} q(\hat{\boldsymbol{y}}_n|s)(\hat{\boldsymbol{y}}_n-\boldsymbol{\mu}(\mathcal{S}_{\mathrm{f}}))(\hat{\boldsymbol{y}}_n-\boldsymbol{\mu}(\mathcal{S}_{\mathrm{f}}))^{\top}\right\}\\
&=\mathrm{Tr}\left\{(\boldsymbol{\Sigma}(\mathcal{S}_{\mathrm{f}}))^{-1}(\boldsymbol{\Sigma}(\mathcal{S}_{\mathrm{f}}))\phi(\mathcal{S}_{\mathrm{f}})\right\}=L\phi(\mathcal{S}_{\mathrm{f}}) \qquad (15)
\end{aligned}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace of a square matrix. Then, Eq. (11) can be rewritten as:

$$\gamma(\mathcal{S}_{\mathrm{f}})=-\frac{1}{2}[L\ln(2\pi)+\ln|\boldsymbol{\Sigma}(\mathcal{S}_{\mathrm{f}}))|+L]\phi(\mathcal{S}_{\mathrm{f}}) \qquad (16)$$

Thus, the log-likelihood $\gamma(\mathcal{S}_{\mathrm{f}})$ depends only on the pooled state occupancy $\phi(\mathcal{S}_{\mathrm{f}})$ and the pooled state variance $\boldsymbol{\Sigma}(\mathcal{S}_{\mathrm{f}})$. Both can be calculated from the saved parameters of state occupancy counts, means, and variances for all HMM states during the preceding Baum–Welch re-estimation.

By utilizing the question chosen from an automatically generated question set, the node associated with $\mathcal{S}_{\mathrm{f}}$ is divided into two child nodes: a left node containing subset $\mathcal{S}$l and a right node containing subset $\mathcal{S}_{\mathrm{r}}$. This splitting operation aims to maximize the increase in log-likelihood, which can be computed as follows:

$$\Delta\gamma=\gamma(\mathcal{S}_{\mathrm{l}})+\gamma(\mathcal{S}_{\mathrm{r}})-\gamma(\mathcal{S}_{\mathrm{f}}) \qquad (17)$$

where $\gamma(\mathcal{S}_{\mathrm{l}})$ and $\gamma(\mathcal{S}_{\mathrm{r}})$ are the log-likelihoods of the state set in the left child node and the right child node, respectively.

To generate the question set, we start by placing all feature frames within the root node of a binary decision tree. Subsequently, we utilize a k-means algorithm ($k=2$) to establish an optimal partition. This process mirrors the binary decision tree for state tying discussed earlier and enhances the loglikelihood of frames by assuming a single Gaussian distribution. This top-down procedure continues until each node exclusively represents a single viseme class. A question can be derived from a nonleaf node by considering all reachable leaves stemming from that node. These questions constitute our question set for state-tying.

IV. Hybrid and Collaborative Viseme Modeling

This section delineates our proposed HVSEM and CVSEM frameworks. We first analyze VSM and E2E error prediction and uncover distinct trends. In particular, for pose-invariant samples, the E2E model outperforms the VSM in terms of accuracy. Conversely, in scenarios where the head poses changes, VSM outperforms E2E modeling. We hypothesize that the E2E model can capture long-term global lip movements, leading to more accurate classification, especially in pose-invariant scenarios. In contrast, VSM is better tuned to analyze motion between adjacent frames, effectively reducing the impact of head motion on recognition performance. We conducted extensive quantitative analyses to support these observations with experimental results, which will be detailed later.

Inspired by the complementary nature of E2E and VSM concerning speaker head movements, we initially introduced HVSEM, which leverages the strengths of both frames through multitask learning (MTL). Subsequently, we present CVSEM, which amplifies the synergy between tasks via an SMTM module. The specifics are described below.

A. HVSEM

As illustrated in Fig. 3, where both the green and black arrows are activated, HVSEM comprises a shared visual front end followed by a shared sequence backend. Additionally, it includes E2E and VSM branches. When provided with a sequence of lip frames denoted as $\boldsymbol{X}$, the visual frontend $\mathcal{F}_V$ extracts framewise visual embeddings $\boldsymbol{E}$, as described by Eq. (4). Subsequently, the sequence back end $\mathcal{F}_S$ aggregates temporal information and generates the hidden representations $\boldsymbol{Y}$, as outlined in Eq. (5).

The E2E branch then aggregates information from all time steps, yielding the representation $\boldsymbol{g}$. Specifically, we utilize the multihead attention mechanism outlined in [15]:

$$\boldsymbol{a}_n=\mathrm{SoftMax}\left(\frac{\sum_{t=1}^{T}\boldsymbol{y}_t\boldsymbol{W}_n^{\mathrm{Q}}(\boldsymbol{Y}\boldsymbol{W}_n^{\mathrm{K}})^{\top}}{T\sqrt{d_k}}\right) \qquad (18)$$

$$\boldsymbol{g}=\mathrm{Concat}(\boldsymbol{a}_1(\boldsymbol{Y}\boldsymbol{W}_1^{\mathrm{V}}),\cdots,\boldsymbol{a}_{N_{\mathrm{H}}}(\boldsymbol{Y}\boldsymbol{W}_{N_{\mathrm{H}}}^{\mathrm{V}}))\boldsymbol{W}_n^{\mathrm{O}}+\frac{\sum_{t=1}^{T}\boldsymbol{y}_t}{T} \qquad (19)$$

where $W_n^Q\in\mathbb{R}^{L\times d_k}$, $W_n^K\in\mathbb{R}^{L\times d_k}$, $W_n^V\in\mathbb{R}^{L\times d_v}$ and $W^O\in\mathbb{R}^{N_{\mathrm{H}}d_v\times L}$ are the projection matrices for the n-th head, and $\boldsymbol{a}_n\in\mathbb{R}^T$ is the attention weight. We employ $N_{\mathrm{H}}=8$ and $d_k=d_v=64$ by default. Finally, $\boldsymbol{g}$ is sent to the linear classifier $\mathcal{F}_{\mathrm{L}}^{\mathrm{word}}$ for word prediction $p^{\mathrm{E2E}}(z|\boldsymbol{X})$:

$$p(z|\boldsymbol{X})=\mathcal{F}_{\mathrm{L}}^{\mathrm{word}}(\boldsymbol{g}) \qquad (20)$$

where $\mathcal{F}_{\mathrm{L}}^{\mathrm{word}}$ is analogous to $\mathcal{F}_{\mathrm{L}}^{\mathrm{state}}$. However, the output dimension of the fully connected layer for $\mathcal{F}_{\mathrm{L}}^{\mathrm{word}}$ is equivalent to N_{W}. On the other hand, the VSM branch directly uses another linear classifier to estimate the state posterior probability $p(s_t|\boldsymbol{x}_t)$, followed by Eq. (6).

In the testing stage, $p(s_t|\boldsymbol{x}_t)$ undergoes decoding to obtain the word posterior probability $p^{\mathrm{VSM}}(z|\boldsymbol{X})$ utilizing the operations described in Eq. (1) and (3). Subsequently, the final posterior probability $p^{\mathrm{final}}(z|\boldsymbol{X})$ is generated by a weighty combination of the word posterior probabilities from both branches:

$$p^{\mathrm{final}}(z|\boldsymbol{X})=\alpha p^{\mathrm{E2E}}(z|\boldsymbol{X})+(1-\alpha)p^{\mathrm{VSM}}(z|\boldsymbol{X}) \qquad (21)$$

where α regulates the contribution of each branch.

The whole model is trained end-to-end via multitask learning. The loss function includes a weighted combination of segment-level and frame-level cross-entropy losses, effectively

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3390148

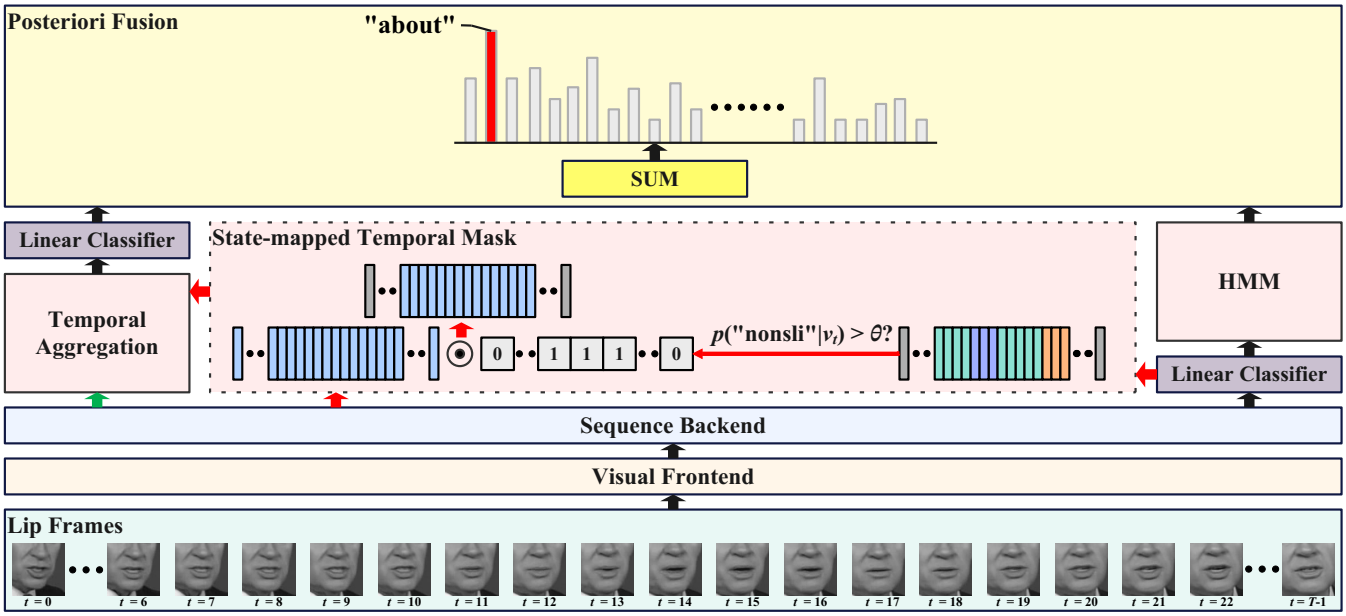


Fig. 3. Illustration of the proposed HVSEM and CVSEM. The key distinction lies in integrating the SMTM module to connect the E2E and VSM branches. HVSEM: Indicated by activating green and black arrows. CVSEM: Indicated by activating red and black arrows.

measuring the multiscale disparity between the predicted and ground-truth text labels:

$$\mathcal{L}=-\sum_{j=0}^{N_{\mathrm{B}}-1}\left[\alpha\log p(z^j|\boldsymbol{X}^j)+(1-\alpha)\sum_{t=0}^{T^j-1}\log p(s_t^j|\boldsymbol{x}_t^j)\right] \quad (22)$$

To maintain consistency between training and testing, α is strategically employed to achieve a balanced integration of the various loss terms.

B. CVSEM

A word and its corresponding state sequence convey identical textual content. HVSEM directly predicts frame-level state and segment-level word outputs from the same input in parallel but fails to capture interactions between words and states. To further exploit these hierarchical interactions, we propose a novel collaborative viseme subword and end-to-end modeling (CVSEM) framework. As illustrated in Fig. 3, when both the red and black arrows are activated, an SMTM module is used to integrate the E2E and VSM branches. The SMTM attempts to selectively filter out unimportant or noisy hidden representations, thereby compelling the E2E branch to make decisions based on the most crucial time steps. Specifically, the SMTM takes frame-level state posterior probabilities from the VSM branch as inputs and outputs a Boolean mask $\boldsymbol{m}=[m_1,\cdots,m_t,\cdots,m_T]$ to the E2E branch:

$$m_t=\left\lceil\sum_{s\in\mathcal{S}_{\mathrm{nonsil}}}p(s|\boldsymbol{x}_t)-\theta\right\rceil \quad (23)$$

where $\mathcal{S}_{\mathrm{nonsil}}$ denotes the set of HMM states corresponding to the nonsilent CD-viseme subwords and $p(s|\boldsymbol{x}_t)$ is estimated by the VSM branch. θ serves as a threshold value that regulates the level of stringency when filtering visual features and plays a critical role in balancing the inclusion of informative features with the exclusion of irrelevant noise. Its value spans from 0 to 1 and is set to 0.5 as a balanced default, providing an optimal compromise between sensitivity and specificity.

Then, $\boldsymbol{m}$ was applied to the raw hidden representation $\boldsymbol{Y}$:

$$\widetilde{\boldsymbol{Y}}=\boldsymbol{m}\odot\boldsymbol{Y} \quad (24)$$

where $\odot$ represents elementwise multiplication and $\widetilde{\boldsymbol{Y}}$ denotes the filtered hidden representation that replaced $\boldsymbol{Y}$ in participating in the subsequent forward process of the E2E branch, as depicted in Eqs. (18), (19) and (5).

TABLE II
THE STATISTICS OF THE LRW AND LRW-1000.

Statistic	LRW	LRW-1000
Words	500	1000
Training instances	488766 (157h)	603190(45h)
Development instances	25000 (8h)	63238(6h)
Test instances	25000(8h)	51590(6h)
Training instances per word	800 ∼ 1000	**25 ∼ 58028**
Language	English	Mandarin
Environment	TV	TV
Resolution	256 × 256	1920 × 1080 or 1024 × 576
Pose	Natural	Natural
Speakers	>1000	>2000

V. EXPERIMENTS AND RESULTS ANALYSIS

We initially employ the same model backbone as described in [15] along with two extensive word-level lip reading datasets: LRW [11] and LRW-1000 [12]. The visual front end, HP-ResNet18, is a modified version of ResNet18. Within HP-ResNet18, the conventional 2D convolution is transformed into

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3390148

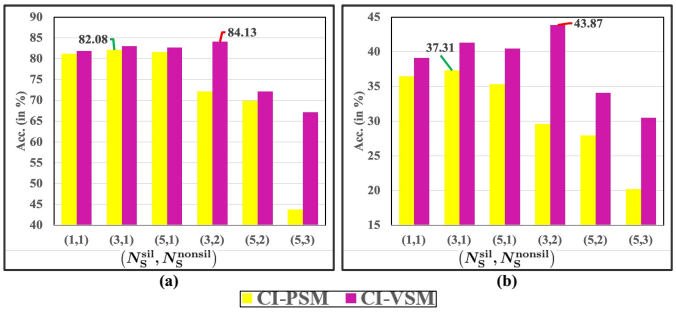


Fig. 4. Comparison of word accuracies (in %) among CI-VSM and CI-PSM with different HMM hyperparameters on the (a) LRW and (b) LRW-1000 datasets. N_S^{nonsil}: the number of HMM states corresponding to nonsilent subwords, N_S^{sil}: the number of HMM states corresponding to silence.

a hierarchical pyramidal convolution (HPConv). A HPConv is characterized by the use of an array of spatial kernels, allowing it to capture both local and global contextual information. It also incorporates a hierarchical connection that facilitates the use of local spatial features as inputs and outputs for extracting global spatial features. The sequence back end is implemented using an MS-TCN [37], which consists of multiple branches with different kernel sizes. The outputs of these branches are concatenated to ensure that each convolution layer integrates information across various temporal scales.

The data statistics are detailed in Table II. Notably, LRW-1000 is highly skewed, with training instances ranging from 25 for the least frequent word to $58,028$ for the most common word. This highlights the challenge posed by the LRW-1000 dataset. To ensure that all the videos have the same number of frames, a practical setting on the LRW-1000 is that 40 continuous frames are selected for each word, and the target word is placed at the center to make it similar to the LRW data. Unless otherwise specified, these datasets are used with only word labels without additional annotations.

In all experiments, all lip-centered videos of size 96×96 pixels are cropped based on the facial landmark points and then converted to grayscale. We employ the same video data augmentations proposed in [15], which include horizontal flipping and random shifts. During training, each video is randomly cropped to 88×88 pixels. We use the Adam [60] optimizer with an initial learning rate of 0.0003 and a weight decay of $10e-4$. Utilizing the standard cosine scheduler [61], we set the total number of epochs to 80 and progressively reduce the learning rate to $10e-7$. Furthermore, a dropout rate of 0.2 is applied to all the models.

A. Performance Analysis of CI and CD VSMs

To highlight our first contribution presented in Section I, we first compare word accuracies (in %) among CI-VSMs using various HMM hyperparameters (i.e., N_S^{nonsil} and N_S^{sil}) on both the LRW and LRW-1000 datasets. We systematically vary the values of N_S^{nosil} from the set $\{1,2,3\}$ and N_S^{sil} from the set $\{1,3,5\}$. Given that "silence" encompasses not only actual pauses but also out-of-vocabulary pronunciations, we typically set $N_S^{\text{nonsil}} < N_S^{\text{sil}}$. This leads to a total of 6 experimental configurations with the bar chart results illustrated in Fig. 4.

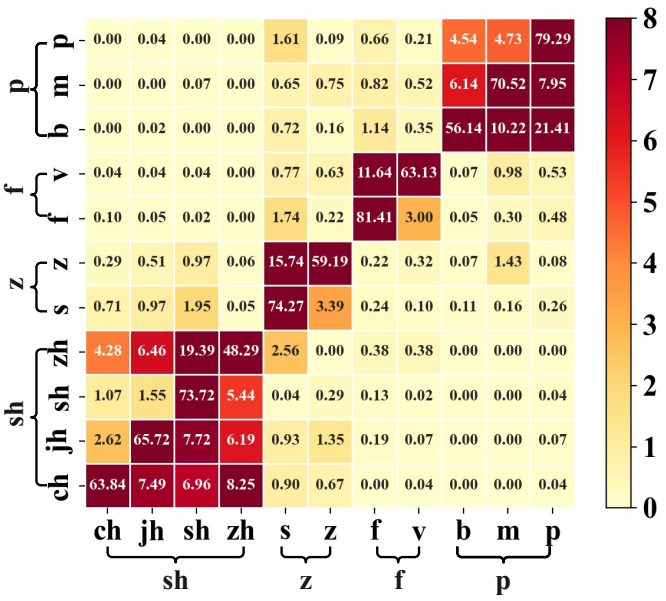
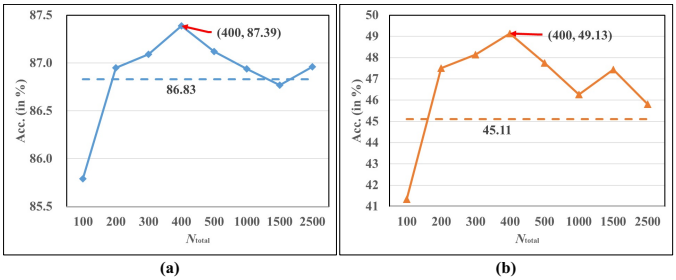


Fig. 5. The confusion matrix of the frame-level phone classification in the LRW dataset is presented in a summarized format, showing the results for 11 phones. Additionally, the corresponding 4 visemes are also indicated.

We observe that CI-VSM attains an optimal configuration with $N_S^{\text{nonsil}} = 2$ and $N_S^{\text{sil}} = 3$ in both the LRW and LRW-1000 datasets. Specifically, on the LRW set, this yields a word accuracy of 84.13% (in Fig. 4(a)), and on the LRW-1000 set, the accuracy is 43.87% (in Fig. 4(b)). This differs from the configuration of $N_S^{\text{nonsil}} = 3$ and $N_S^{\text{sil}} = 5$ commonly used in automatic speech recognition (ASR). A plausible explanation lies in the temporal resolution disparities between acoustic and visual features. The acoustic feature typically boasts a higher temporal resolution, demonstrated by FBANK features with a frame shift of 10 ms, resulting in 100 frames per second. In contrast, videos operate at a lower temporal resolution, capturing only 25 frames per second. A lower temporal resolution results in fewer state occupancies, increasing the suitability of HMMs with fewer states and ultimately leading to improved performance.

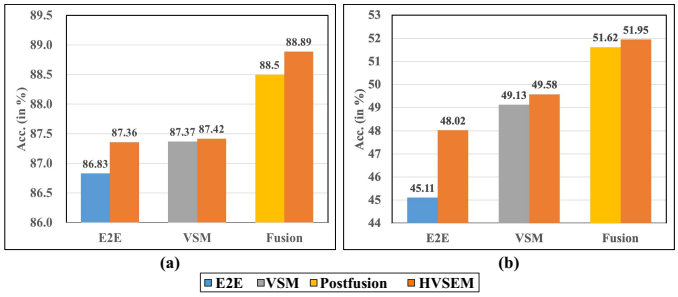
Furthermore, we present the word accuracies (in %) of the context-independent phone subword modeling (CI-PSM) results on the LRW and LRW-1000 datasets in yellow bars in Fig. 4 side-by-side. It is evident that CI-VSM, shown as purple bars, consistently outperforms CI-PSM across all configurations. This highlights the advantages of using visemes due to the strong correlation between visemes and lip shape. Additionally, we compute a confusion matrix for frame-level phone classification using the LRW test set, as shown in Fig. 5. The accuracies for 11 phones, including "ch", "jh", "sh", "zh", "s", "z", "f", "v", "b", "m" and "p", are grouped into 4 visemes, i.e., "sh", "z", "f" and "p". A critical observation is that there is often confusion between phones within the same viseme group. This indicates that forcing the model to distinguish between visually indistinguishable phones can easily lead to errors. Conversely, choosing appropriate visemes as basic units further improves performance.

Even with precise hyperparameter adjustments, CI-VSM



(a) (b)

Fig. 6. Comparison of word accuracies (in %) between the E2E baseline and CD-VSM systems with varying numbers of tied states on the LRW (a) and LRW-1000 (b) datasets. The dotted line shows the results of the E2E baseline.



(a) (b)

Fig. 7. Comparison of word accuracies (in %) among the E2E, VSM, postfusion, and HVSEM systems on the (a) LRW and (b) LRW-1000 datasets. Notably, the HVSEM system generated three results: one from the E2E branch, one from the VSM branch, and the post-fused result.

TABLE III
COMPARISON OF WORD ACCURACIES (IN %) BETWEEN POST-FUSION SYSTEMS WITH DIFFERENT α ON THE LRW AND LRW-1000 DATASETS.

α	0	0.1	0.3	0.5	0.7	0.9	1
LRW	86.83	87.58	87.98	**88.50**	87.99	87.16	87.39
LRW-1000	45.11	49.66	50.59	**51.62**	49.12	46.64	49.13

TABLE IV
COMPARISON OF WORD ACCURACIES (IN %) AMONG E2E, VSM AND POSTFUSION SYSTEMS ON DIFFERENT SUBSETS WITH THE SAME NUMBER OF HEAD MOVEMENTS ON THE LRW DATASET.

Movements	**0**	**1**	**2**	**≥ 3**
E2E	**90.43**	85.12	77.21	70.52
VSM	87.72	**87.21**	**86.73**	**85.52**
Postfusion ($\alpha = 0.5$)	90.08	86.59	84.83	83.75

still yields lower word accuracies than the E2E baselines (86.83% and 45.11%), prompting our focus toward CD-VSM. We compare the accuracies of the E2E and CD-VSM systems for various numbers of tied states on the LRW and LRW-1000 datasets. As depicted in Fig. 6, as the number of states decreases, the accuracies on both the LRW and LRW-1000 datasets initially increase before gradually decreasing. The highest accuracies, reaching 87.39% for the LRW dataset and 49.13% for the LRW-1000 dataset, are achieved when $N_{\text{S}} = 400$.

These results represent a significant improvement over CI-VSM, showing absolute accuracy gains of 3.26% and 5.26% on the LRW and LRW-1000 datasets, respectively. This highlights the critical role of explicitly modeling the coarticulation effect on lip movements to achieve greater accuracy. Additionally, state tying balances the training data distribution, ensuring sufficient data for each state and further improving accuracy. Moreover, the CD-VSM results also show significant superiority over the E2E baseline, achieving relative error rate reductions of 4.25% and 7.32% on the LRW and LRW-1000 datasets, respectively. This finding underscores that VSM provides dual advantages in both performance and interpretability.

B. Verification of complementarity between VSM and E2E

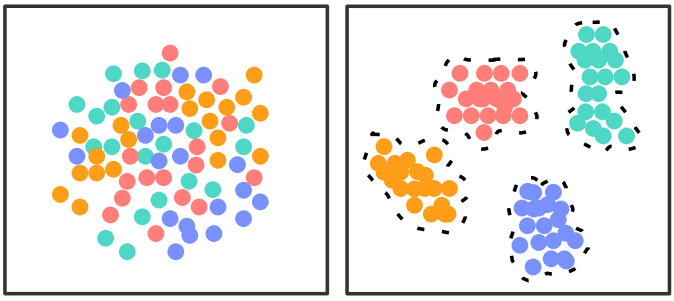
To show our second contribution presented in Section I, a straightforward posterior fusion is employed to validate the complementarity between the VSM and E2E models. The CD-VSM and E2E systems are trained individually to generate word posterior probabilities and the final decision is made using a weighted average as illustrated in Eq. (21). Table III lists the word accuracies (in %) of postfusion systems employing various weight factors denoted as α. We choose an α value from the set: $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. Notably, postfusion achieves peak accuracy at $\alpha = 0.5$ for both LRW and LRW-1000, outperforming the best single system with absolute improvements of 1.11% and 2.49%, respectively. This finding underscores the superior performance of the postfusion system by exploiting complementarity and highlights its robustness across different hyperparameter values. We then employ default values of $\alpha = 0.5$ in all follow-up experiments.

To further clarify the complementarity, we employed HopeNet [62] to estimate the head pose of the target speaker in the video. A change in intrinsic Euler angles (yaw, pitch, and roll) greater than 30 degrees was considered to indicate the occurrence of a single head movement. We then categorized the LRW test set into four subsets, 0, 1, 2 and ≥ 3 based on the frequency of head movements. The word accuracies (in %) of the E2E, VSM, and postfusion systems for these subsets are presented in Table IV. As head movement increases, E2E performance decreases. VSM remains relatively robust performance but falls behind in head-invariant samples. The postfusion system excels by exploiting the complementarity of head movement. A similar trend can be seen in the LRW-1000 set, but page limitations do not permit a detailed discussion.

C. Performance analysis of HVSEM

To verify the effectiveness of our second contribution presented in Section I with the proposed HVSEM approach, we present a comparison of word accuracies among the E2E, VSM, postfusion, and HVSEM systems on the LRW and LRW-1000 datasets, with bar charts shown in Fig. 7. We find that HVSEM achieves the highest word accuracies. However, the improvement over the postfusion system is not very significant for either the LRW or the LRW-1000 datasets. It is worth noting that the number of parameters is now halved compared to that of the postfusion system.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2024.3390148



(a) (b)

Fig. 8. Frame-level hidden representation visualization of the (a) E2E and (b) HVSEM systems. Note that the same color indicates hidden representations belonging to the same HMM state.

TABLE V
COMPARISON OF WORD ACCURACIES (IN %) AMONG HVSEM AND CVSEM SYSTEMS. WE ADDITIONALLY RETRAINED THE HVSEM SYSTEMS USING THE GROUND-TRUTH WORD BOUNDARIES AND THE WORD BOUNDARIES DERIVED THROUGH FORCE ALIGNMENT, DENOTED AS HVSEM+WB AND HVSEM+FA, RESPECTIVELY.

Dataset	LRW			LRW-1000		
Branch	E2E	VSM	Fusion	E2E	VSM	Fusion
HVSEM	87.36	87.42	88.89	48.02	49.58	51.95
CVSEM	**88.35**	87.45	**89.40**	**50.82**	49.87	**54.31**
HVSEM+WB	**89.84**	87.53	**90.66**	**53.85**	49.92	**56.17**
HVSEM+FA	**90.44**	87.58	**91.08**	**55.28**	50.03	**57.05**

To further analyze the improvements achieved by HVSEM, we present the results of both the E2E and VSM branches in Fig. 7. The E2E branch primarily contributes to the final improvement. Specifically, compared to the single E2E system, the E2E branch achieves relative error rate reductions of 4.19% and 5.30% on the LRW and LRW-1000 datasets, respectively. However, the improvement in the VSM branch is not as pronounced. Therefore, we propose that supervision from frame-level state labels enables the sequence back end to generate more discriminative frame-level hidden representations, which is beneficial for accurate judgments. To validate our hypothesis, we utilized t-SNE [63] to visualize the hidden representations generated by the E2E and HVSEM systems with random samples from the validation set. As shown in Fig. 8, the hidden representations in the E2E system appear to be scattered across different state categories. In contrast, the HVSEM system displays significantly more discriminative hidden representations.

D. Performance analysis of CVSEM

To verify the effectiveness of our third contribution presented in Section I using CVSEM, we present a comparative analysis of word accuracies (in %) between HVSEM and CVSEM, focusing on the outcomes in Fig.3 of the E2E and VSM branches and the fusion results. As illustrated in Table V, CVSEM outperforms HVSEM in all the other scenarios, achieving relative error rate reductions of 4.59% and 4.91% on the LRW and LRW-1000 datasets, respectively. This

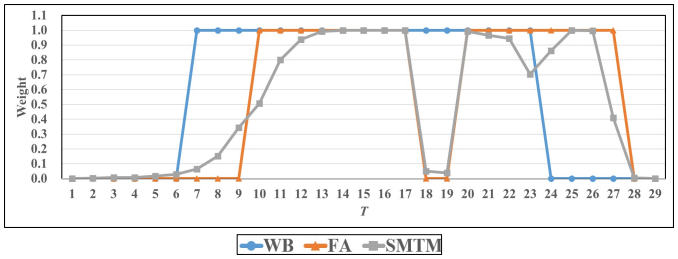


Fig. 9. Visualization of word boundaries from a randomly sampled example in the LRW dataset. The blue, orange and gray lines represent the ground truth, the force alignment and the SMTM module, respectively. Note that for the SMTM module, the raw probability values are depicted.

notable improvement can be primarily attributed to the SMTM module, which effectively filters out frames unrelated to the target word based on the frame-level state posterior probability, resulting in substantial performance enhancements for the E2E branch, with relative error rate reductions of 19.62% and 11.22% on the LRW and LRW-1000 tasks, respectively.

To further explore the efficacy of SMTM, we retrained the HVSEM systems using ground-truth word boundaries and those derived through force alignment, denoted HVSEM+WB and HVSEM+FA, respectively. The word accuracies of HVSEM+WB and HVSEM+FA are also presented in Table V. Incorporating word boundaries significantly enhances the performance of the E2E branch, resulting in a substantial improvement in the fusion results. Interestingly, the HVSEM+FA system yields higher word accuracy than does the HVSEM+WB system. Specifically, by using force alignment-derived word boundaries, the E2E branch achieves 5.90% and 3.10% decreases in relative error rates.

In Fig. 9, we visually compare the ground-truth word boundaries, those derived through force alignment, and the outcomes from the SMTM module. Upon meticulous examination of each lip frame, it becomes apparent that the word boundaries acquired via force alignment not only offer more accurate delineation of both frames but also indicate pauses during articulation. Consequently, with the guidance of frame-level state labels, the SMTM model can allocate weights to individual frames, thereby mitigating the impact of frame-level noise and resulting in discernible improvement.

E. Robustness toward powerful model backbones

We assess the robustness of three proposed system frameworks (VSM, HVSEM and CVSEM) toward different model backbones using the LRW and LRW-1000 sets. For the visual front end, we incorporate the time shift module (TSM) into the preexisting HP-ResNet18, as inspired by [26]. This modified architecture is denoted as HP-ResNet18-TSM, where we specifically adopt a 1/4 proportion residual shift. For the sequence back end, we employ the DenseTCN as described in [38]. This network is renowned for extracting temporal features at denser scales, thus reinforcing feature robustness without compromising the maximum receptive field size. We adopt partially dense blocks with hyperparameters set at their reported optimum. We replaced the corresponding parts in the

TABLE VI
COMPARISON OF WORD ACCURACY (IN %) AMONG E2E, VSM, HVSEM AND CVSEM FRAMEWORKS WITH POWERFUL MODEL BACKBONES ON THE LRW AND LRW-1000 DATASETS.

Model	Framework	LRW	LRW-1000
HP-ResNet18 +MS-TCN	E2E	86.83	45.11
	VSM	87.39	49.13
	HVSEM	88.89	51.95
	CVSEM	**89.40**	**54.31**
HP-ResNet18-TSM +MS-TCN	E2E	87.74	47.62
	VSM	88.17	51.66
	HVSEM	89.67	54.14
	CVSEM	**90.08**	**56.28**
HP-ResNet18 +DenseTCN	E2E	87.97	48.58
	VSM	88.43	52.18
	HVSEM	89.86	55.02
	CVSEM	**90.36**	**57.40**
HP-ResNet18-TSM +DenseTCN	E2E	88.49	50.48
	VSM	89.05	54.64
	HVSEM	90.24	56.87
	CVSEM	**90.75**	**58.89**

baseline model with HP-ResNet18-TSM and DenseTCN, and the accuracies (in %) are presented in Table VI.

As evident from the tabulated results, within the E2E framework, HP-ResNet18-TSM and DenseTCN individually achieved good absolute accuracy improvements of 0.91% and 1.14% on the LRW set compared to the baseline result of 86.83%. When combined, the improvement goes to 88.49% with an increased accuracy of 1.66% shown in the bottom block, indicating a synergistic effect. The improvements are even more pronounced on the LRW-1000 dataset, with gains of 2.51%, 3.47% and 5.37% achieved by HP-ResNet18-TSM, DenseTCN and their combination, respectively. This highlights the effectiveness of these model backbones.

Across all back ends, VSM consistently outperforms E2E modeling. The proposed HVSEM and CVSEM frameworks exhibit remarkable performance improvements, with CVSEM outperforming HVSEM. Even with the most powerful model backbone, VSM achieves impressive absolute accuracy improvements of 0.56% and 4.16% on the LRW and LRW-1000 datasets, respectively, over E2E modeling. Compared with VSM, HVSEM yields significant improvements, with absolute accuracy gains of 1.19% and 2.23%. Similarly, CVSEM yields significant improvements in the absolute accuracy of 0.51% and 2.02% over HVSEM. This indicates that the proposed VSM, HVSEM, and CVSEM frameworks are transferable and adaptable to various model backbones.

In Table VII, we present the performance comparison between our approach and the prior SOTA methods on both the LRW and LRW-1000 datasets. On the LRW dataset, our method achieves an accuracy of 90.75%, establishing a new SOTA model with an absolute improvement of 1.23% over the current SOTA model. Moreover, on the LRW-1000 dataset, our method achieves an even higher accuracy of 58.89%, surpassing the best baseline method, which achieved 56.85%. This further underscores the effectiveness and generalizability of our proposed method.

TABLE VII
COMPARISON OF WORD ACCURACY (IN %) BETWEEN OUR WORK AND SOTA METHODS ON THE LRW AND LRW-1000 DATASETS. NONE OF THE RESULTS LISTED USES EXTRA ANNOTATIONS, SUCH AS WORD BOUNDARIES.

Method	LRW	LRW-1000
ResNet34+BiLSTM [10]	83	\
Two-Stream ResNet18+BiLSTM [31]	84.1	\
ResNet18+STFM [13]	83.7	\
Two-Stream ResNet18+BiGRU [32]	84.1	41.9
ResNet34+GCN+BiGRU [33]	84.25	\
P3D-50+BiGRU [29]	84.48	\
ResNet18+MS-TCN [37]	85.3	41.4
ResNet18+Transformer [44]	87.32	56.85
ResNet18+ASST-GCN+MS-TCN [34]	85.5	\
SE-ResNet18+BiGRU [14]	86.2	48.3
ResNet18+TSM+BiGRU [26]	86.23	44.6
HP-ResNet18+MS-TCN+Attention [15]	86.83	45.11
ALSOS-ResNet18+MS-TCN [27]	87.01	\
ResNet18+MS-TCN+MS-KD [39]	87.9	\
ResNet18+DenseTCN [38]	88.34	43.65
SE-ResNet18+BiGRU+VTM [43]	86.9	49.7
ResNet18+MS-TCN+MVM [46]	88.5	53.82
EfficientNetV2+Transformer+TCN [28]	89.52	\
HP-ResNet18-TSM+DenseTCN+CVSEM	**90.75**	**58.89**

VI. CONCLUSION

In this study, we proposed three frameworks, namely, VSM, HVSEM and CVSEM, for DNN-based word-level lip reading. In VSM, a word is decomposed into a sequence of viseme subwords. DNNs and HMMs establish associations between each lip frame and the viseme subword, culminating in the final word decision. HVSEM exploits the complementarity of the E2E and VSM frameworks with respect to speaker head movements via MTL. CVSEM further enhances task synergy by integrating an SMTM module with HVSEM. Experiments with different model backbones show that CD-VSMs with tied states exhibit superior performance compared to E2E modelings. In addition, the HVSEM method consistently outperforms VSM, highlighting the effectiveness of joint modeling through MTL. Furthermore, CVSEM consistently outperforms HVSEM, highlighting the effectiveness of filtering irrelevant frames with SMTM. In the future, we plan to extend the proposed framework to sentence-level lip reading.

VII. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China under Grant No.62171427.

REFERENCES

[1] A. Haliassos, K. Vougioukas, S. Petridis *et al.*, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proc. CVPR 2021*, 2021, pp. 5039–5049.

[2] K. Sun, C. Yu, W. Shi *et al.*, "Lip-interact: Improving mobile device interaction with silent speech commands," in *Proc. UIST 2018*, 2018.

[3] N. Tye-Murray, M. S. Sommers *et al.*, "Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing," *Ear and hearing*, vol. 28, no. 5, pp. 656–668, 2007.

[4] H. Chen, H. Zhou, J. Du *et al.*, "The first multimodal information based speech processing (misp) challenge: Data, tasks, baselines and results," in *Proc. ICASSP 2022*, 2022.

[5] H. Chen, J. Du, Y. Dai *et al.*, "Audio-visual speech recognition in misp2021 challenge: Dataset release and deep analysis," in *Proc. INTERSPEECH 2022*, 2022.

[6] H. Chen, S. Wu, Y. Dai *et al.*, "Summary on the multimodal information based speech processing (misp) 2022 challenge," in *Proc. ICASSP*, 2023.

[7] H. Chen, J. Du, Y. Hu *et al.*, "Correlating subword articulation with lip shapes for embedding aware audio-visual speech enhancement," *Neural Networks*, vol. 143, pp. 171–182, 2021.

[8] Z. Zhou, G. Zhao, X. Hong *et al.*, "A review of recent advances in visual speech decoding," *Image and Vision Computing*, vol. 32, no. 9, pp. 590–605, 2014.

[9] J. S. Chung, A. Senior, O. Vinyals *et al.*, "Lip reading sentences in the wild," in *Proc. CVPR 2017*, 2017, pp. 3444–3453.

[10] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with lstms for lipreading," in *Proc. Interspeech 2017*, 2017, pp. 3652–3656.

[11] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. ACCV 2016*, 2016.

[12] S. Yang, Y. Zhang, D. Feng *et al.*, "Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *Proc. FG 2019*, 2019, pp. 1–8.

[13] X. Zhang, F. Cheng, and W. Shilin, "Spatio-temporal fusion based convolutional sequence learning for lip reading," in *Proc. ICCV 2019*, 2019, pp. 713–722.

[14] D. Feng, S. Yang, and S. Shan, "An efficient software for building lip reading models without pains," in *Proc. ICMEW 2021*, 2021, pp. 1–2.

[15] H. Chen, J. Du, Y. Hu *et al.*, "Automatic lip-reading with hierarchical pyramidal convolution and self-attention for image sequences with no word boundaries," in *Proc. INTERSPEECH 2021*, 2021, pp. 3001–3005.

[16] T. Stafylakis, M. H. Khan, and G. Tzimiropoulos, "Pushing the boundaries of audiovisual word recognition using residual networks and lstms," *Computer Vision and Image Understanding*, vol. 176, pp. 22–32, 2018.

[17] K. Livescu, E. Fosler-Lussier, and F. Metze, "Subword modeling for automatic speech recognition: Past, present, and emerging approaches," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 44–57, 2012.

[18] C. Neti, G. Potamianos, J. Luettin *et al.*, "Audio-visual speech recognition," in *Workshop 2000 Final Report*, 2000, pp. 1–86.

[19] P. Lucey, G. Potamianos, and S. Sridharan, "A unified approach to multi-pose audio-visual asr," in *Proc. Interspeech 2007*, 2007, pp. 650–653.

[20] G. Zhao, M. Barnard, and M. Pietikainen, "Lipreading with local spatiotemporal descriptors," *IEEE Transactions on Multimedia*, vol. 11, no. 7, pp. 1254–1265, 2009.

[21] H. E. Cetingul, Y. Yemez, E. Erzin *et al.*, "Discriminative analysis of lip motion features for speaker identification and speech-reading," *IEEE Transactions on Image Processing*, vol. 15, no. 10, pp. 2879–2891, 2006.

[22] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

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

[24] P. Wu, H. Liu, X. Li *et al.*, "A novel lip descriptor for audio-visual keyword spotting based on adaptive decision fusion," *IEEE Transactions on Multimedia*, vol. 18, no. 3, pp. 326–338, 2016.

[25] S. Petridis, T. Stafylakis, P. Ma *et al.*, "End-to-end audiovisual speech recognition," in *Proc. ICASSP 2018*, 2018, pp. 6548–6552.

[26] M. Hao, M. Mamut, N. Yadikar *et al.*, "How to use time information effectively? combining with time shift module for lipreading," in *Proc. ICASSP 2021*, 2021, pp. 7988–7992.

[27] D. Tsourounis, D. Kastaniotis, and S. Fotopoulos, "Lip reading by alternating between spatiotemporal and spatial convolutions," *Journal of Imaging*, vol. 7, no. 5, p. 91, 2021.

[28] A. Koumparoulis and G. Potamianos, "Accurate and resource-efficient lipreading with efficientnetv2 and transformers," in *Proc. ICASSP 2022*, 2022, pp. 8467–8471.

[29] B. Xu, C. Lu, Y. Guo *et al.*, "Discriminative multi-modality speech recognition," in *Proc. CVPR 2020*, 2020, pp. 14 433–14 442.

[30] C.-H. Wang, "Multi-grained spatio-temporal modeling for lip-reading," in *Proc. BMVC 2019*, 2019, pp. 1–11.

[31] X. S. Weng and K. Kitani, "Learning spatiotemporal features with two-stream deep 3d cnns for lipreading," in *Proc. BMVC*, 2019, pp. 1–13.

[32] J. Xiao, S. Yang, Y. Zhang *et al.*, "Deformation flow based two-stream network for lip reading," in *Proc. FG 2020*, 2020, pp. 364–370.

[33] H. Liu, Z. Chen, and B. Yang, "Lip graph assisted audio-visual speech recognition using bidirectional synchronous fusion," in *Proc. INTERSPEECH 2020*, 2020, pp. 3520–3524.

[34] C. Sheng, X. Zhu, H. Xu *et al.*, "Adaptive semantic-spatio-temporal graph convolutional network for lip reading," *IEEE Transactions on Multimedia*, vol. 24, pp. 3545–3557, 2021.

[35] M. Luo, S. Yang, S. Shan *et al.*, "Pseudo-convolutional policy gradient for sequence-to-sequence lip-reading," in *Proc. FG*, 2020, pp. 273–280.

[36] Y. Zhang, S. Yang, J. Xiao *et al.*, "Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition," in *Proc. FG 2020*, 2020, pp. 356–363.

[37] B. Martinez, P. Ma, S. Petridis *et al.*, "Lipreading using temporal convolutional networks," in *Proc. ICASSP 2020*, 2020, pp. 6319–6323.

[38] P. Ma, Y. Wang, J. Shen *et al.*, "Lip-reading with densely connected temporal convolutional networks," in *Proc. WACV*, 2021, pp. 2856–2865.

[39] P. Ma, B. Martinez, S. Petridis *et al.*, "Towards practical lipreading with distilled and efficient models," in *Proc. ICASSP*, 2021, pp. 7608–7612.

[40] S. Ren, Y. Du, J. Lv *et al.*, "Learning from the master: Distilling cross-modal advanced knowledge for lip reading," in *Proc. CVPR 2021*, 2021, pp. 13 320–13 328.

[41] P. Ma, R. Mira, S. Petridis *et al.*, "Lira: Learning visual speech representations from audio through self-supervision," in *Proc. Interspeech 2021*, 2021, pp. 3011–3015.

[42] X. Zhao, S. Yang, S. Shan *et al.*, "Mutual information maximization for effective lip reading," in *Proc. FG 2020*, 2020, pp. 420–427.

[43] C. Sheng, L. Liu, W. Deng *et al.*, "Importance-aware information bottleneck learning paradigm for lip reading," *IEEE Transactions on Multimedia*, pp. 1–13, 2022.

[44] M. Luo, S. Yang, X. Chen *et al.*, "Synchronous bidirectional learning for multilingual lip reading," in *Proc. BMVC 2020*, 2020, pp. 1–13.

[45] M. Kim, J. Hong, S. J. Park *et al.*, "Multi-modality associative bridging through memory: Speech sound recollected from face video," in *Proc. ICCV 2021*, 2021, pp. 296–306.

[46] M. Kim, J. H. Yeo, and Y. M. Ro, "Distinguishing homophenes using multi-head visual-audio memory for lip reading," in *Proc. AAAI 2022*, vol. 22, 2022.

[47] A. Cruttenden, *Gimson's pronunciation of English*. Routledge, 2014.

[48] C. G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech and Hearing Research*, vol. 11, no. 4, pp. 796–804, 1968.

[49] A. A. Montgomery and P. L. Jackson, "Physical characteristics of the lips underlying vowel lipreading performance," *The Journal of the Acoustical Society of America*, vol. 73, no. 6, pp. 2134–2144, 1983.

[50] S. Lee and D. Yook, "Audio-to-visual conversion using hidden markov models," in *Proc. PRICAI 2002*, 2002, pp. 563–570.

[51] M. Gur and D. M. Snodderly, "Direction selectivity in v1 of alert monkeys: evidence for parallel pathways for motion processing." *The Journal of physiology*, vol. 585, pp. 383–400, 2007.

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

[53] S. Ruder, J. Bingel, I. Augenstein *et al.*, "Latent multi-task architecture learning," in *Proc. AAAI 2019*, vol. 33, no. 01, 2019, pp. 4822–4829.

[54] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. CVPR 2018*, 2018, pp. 7482–7491.

[55] Y. Zhang and Q. Yang, "A survey on multi-task learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5586–5609, 2021.

[56] T. Gong, T. Lee, C. Stephenson *et al.*, "A comparison of loss weighting strategies for multi task learning in deep neural networks," *IEEE Access*, vol. 7, pp. 141 627–141 632, 2019.

[57] S. Young, G. Evermann, M. Gales *et al.*, "The htk book," *Cambridge university engineering department*, vol. 3, no. 175, p. 12, 2002.

[58] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[59] D. Povey, A. Ghoshal, G. Boulianne *et al.*, "The kaldi speech recognition toolkit," in *Proc. ASRU 2011*, no. 1, 2011.

[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR 2015*, 2015.

[61] I. Loshchilov and F. Hutter, "SGDR: stochastic gradient descent with warm restarts," in *Proc. ICLR 2017*, 2017.

[62] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. CVPRW 2018*, June 2018.

[63] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

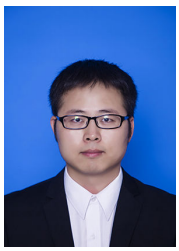
Hang Chen (Student Member, IEEE) received his B.Eng. degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2018. He is currently working toward a Ph.D. degree at the USTC, Hefei, China. His current research includes audio-visual speech enhancement and recognition.

Qing Wang received B.S. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2012 and 2018, respectively. From July 2018 to February 2020, she worked on speech enhancement at Tencent. She is currently a special associate researcher at USTC. Her research interests include speech enhancement, robust speech recognition, audio-visual scene classification, and sound event localization and detection.

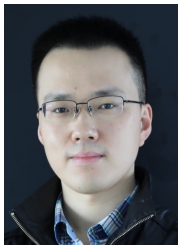
Jun Du (Senior Member, IEEE) received B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. From 2009 to 2010, he was with iFLYTEK Research as a team leader, working on speech recognition. From 2010 to 2013, he joined Microsoft Research Asia as an associate researcher, working on handwriting recognition and OCR. Since 2013, he has been with the National Engineering Laboratory for Speech and Language Information Processing, USTC. He has authored or coauthored more than 150 papers. His main research interests include speech signal processing and pattern recognition applications. He is an associate editor for the IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing and a Member of the IEEE Speech and Language Processing Technical Committee. He was the recipient of the 2018 IEEE Signal Processing Society Best Paper Award. His team won several champions of the CHiME-4/CHiME-5/CHiME-6 Challenge, the SELD Task of the 2020 DCASE Challenge, and the DIHARD-III Challenge.

Genshun Wan received B.S. and M.S. degrees from Jiangsu University, Zhenjiang, China, in 2012 and 2015, respectively. He is currently working toward a Ph.D. degree at the University of Science and Technology of China, Hefei, China. Since 2015, he has been performing iFlytek Research on speech recognition and spoken dialog systems. His general area of research includes speech recognition, dialog systems, and machine learning. He won first place in many speech challenges, including OpenASR Challenge 2021 and the CHiME Challenge 7.

Shifu Xiong received a Bachelor of Science degree from the School of Electronic Information and Communication at Huazhong University of Science and Technology in Wuhan, China, in 2011. Subsequently, in 2014, he obtained a Master of Science degree from the Department of Electronic Engineering and Information Science at the University of Science and Technology of China in Hefei, China. Since 2014, he has been working with iFLYTEK Research on spoken dialog systems. His research interests include speech recognition and natural language processing.



Bao-Cai Yin received B.Eng. and Ph.D. degrees from the University of Science and Technology of China, Hefei, Anhui, China, in 2012 and 2022, respectively. He is currently working as a researcher at iFLYTEK. He joined iFLYTEK in 2015 and has been engaged in artificial intelligence research. His research interests include optical character recognition, medical image analysis and multimodal learning. He won first place in many computer vision challenges, including the Lung Nodule Analysis (LUNA) and the IDRiD Diabetic Retinopathy Segmentation and Grading Challenge.



Jia Pan received B.S., M.S. and Ph.D. degrees in 2006, 2009 and 2021, respectively, from the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China. Since 2009, he has been with iFLYTEK Research on speech recognition and spoken dialog systems. His current research interests include speech recognition and machine learning. He won first place in many speech challenges, including the Open Automatic Speech Recognition (OpenASR) Challenge 2021 and the Blizzard Challenge.

Chin-Hui Lee (Fellow, IEEE) is a professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology. Before joining academia in 2001, he had 20 years of industrial experience, ending at Bell Laboratories, Murray Hill, NJ, USA, as a Distinguished Member of Technical Staff and the Director of the Dialog Systems Research Department. He has authored or coauthored more than 600 papers and 30 patents and has been cited more than 80000 times for his original contributions, with an h-index of 80 on Google Scholar. He has received numerous awards, including the Bell Labs President's Gold Award in 1998. He also won SPS's 2006 Technical Achievement Award for Exceptional Contributions to the Field of Automatic Speech Recognition. In 2012, he was invited by the ICASSP to give a plenary talk on the future of speech recognition. In the same year, he was awarded the ISCA Medal in scientific achievement for pioneering and seminal contributions to the principles and practice of automatic speech and speaker recognition. He is also a Fellow of ISCA.