A Novel Lip Descriptor for Audio-Visual Keyword Spotting Based on Adaptive Decision Fusion

Pingping Wu, Hong Liu, Xiaofei Li, Ting Fan, Xuewu Zhang

A Novel Lip Descriptor for Audio-Visual Keyword Spotting Based on Adaptive Decision Fusion

Pingping Wu, Hong Liu*, Xiaofei Li, Ting Fan, and Xuewu Zhang

***Abstract*—Keyword spotting remains a challenge when applied to real-world environments with dramatically changing noise. In recent studies, audio-visual integration methods have demonstrated superiorities since visual speech is not influenced by acoustic noise. However, for visual speech recognition, individual utterance mannerisms can lead to confusion and false recognition. To solve this problem, a novel lip descriptor is presented involving both geometry-based and appearance-based features in this paper. Specifically, a set of geometry-based features is proposed based on an advanced facial landmark localization method. In order to obtain robust and discriminative representation, a spatiotemporal lip feature is put forward concerning similarities among textons and mapping the feature to intra-class subspace. Moreover, a parallel two-step keyword spotting strategy based on decision fusion is proposed in order to make the best use of audio-visual speech and adapt to diverse noise conditions. Weights generated using a neural network combine acoustic and visual contributions. Experimental results on OuluVS dataset and PKU-AV dataset demonstrate that the proposed lip descriptor shows competitive performance compared to the state of the art. Additionally, the proposed audio-visual keyword spotting method based on decision-level fusion significantly improves the noise robustness and attains better performance than feature-level fusion, which is also capable of adapting to various noisy conditions.**

***Index Terms*—Keyword spotting, Audio-visual fusion, Visual speech recognition, Noisy conditions.**

I. Introduction

AUTOMATIC speech recognition (ASR) has gained wide research attention in the past decades [1]–[3]. In some scenarios, continuous speech recognition that performs a complete transcription is not necessary since the key information lies in only part of the input utterance [4]. Alternatively, keyword spotting (KWS) deals with the identification of some predefined words instead of the whole utterance and can obtain fast access to the key information [5], [6]. Compared with continuous speech recognition, KWS has the capability to cope with situations where various disfluencies and artifacts make the full-scale speech recognition difficult. Besides, without entire utterances to decode, KWS also leads to less time complexity. Therefore, KWS is more suitable for special applications like human-robot interaction (HRI).

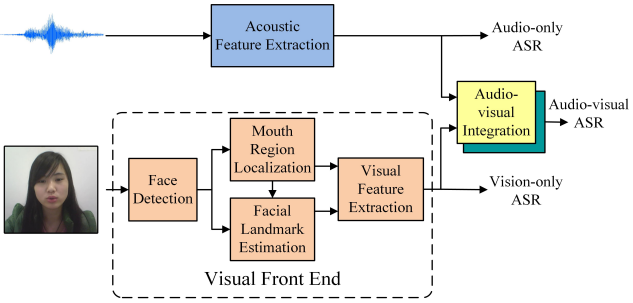


Fig. 1: Primary modules of an AV-ASR system

For KWS, there are three typical approaches: HMM-filler based KWS, phoneme lattice based KWS and large vocabulary continuous speech recognition (LVCSR) based KWS [7]. The most common KWS approach is LVCSR-based KWS which uses an LVCSR system to generate a word lattice and then perform a search within the lattice for the keyword. Although state-of-the-art KWS technology has achieved significant progress and has been successfully applied to some well-defined applications [8]–[10], its performance degrades heavily when applied to real-world environments due to the massive corruption of speech signals.

In order to improve the performance of ASR in the presence of noise, numerous methods have been explored, one of which employs the visual information of vocal organs during the articulating process. Indeed, the intrinsic mechanism of both human speech production and perception is bimodal [11]. When we communicate with others, we not only "listen" but also "look". Moreover, visual information is not affected by the acoustic environment. Therefore, audio-visual automatic speech recognition (AV-ASR) that combines visual speech with acoustic speech, is widely investigated to improve noise robustness [12]–[16]. Fig. 1 shows the basic diagram of AV-ASR, including acoustic feature extraction, visual front end design and audio-visual integration. Visual front end design includes face detection, lip localization and visual feature extraction.

While extensive research has been conducted on AV-ASR, few studies address audio-visual keyword spotting (AV-KWS). Ming Liu *et al.* designed an English-oriented audio-visual word spotter based on feature-level fusion without any adaption to various noisy conditions [17]. Shivappa proposed a hierarchical audio-visual cue integration framework for activity analysis in intelligent meeting rooms where KWS is merely a

Pingping Wu and Hong Liu* (corresponding author) are with the Key Laboratory of Machine Perception (Ministry of Eduction) and the Engineering Lab on Intelligent Perception for Internet of Things (ELIP), Shenzhen Graduate School, Peking University, E-mail: *wupingping@pku.edu.cn*; *hongliu@pku.edu.cn*.

Xiaofei Li is with the PERCEPTION team at INRIA Grenoble Rhône-Alpes, France, E-mail: *xiaofei.li@inria.fr*.

Ting Fan and Xuewu Zhang are with the Engineering Lab on Intelligent Perception for Internet of Things (ELIP), Shenzhen Graduate School, Peking University, E-mail: *fanting*19900126@126.*com*; *zhangxuewu@sz.pku.edu.cn*.

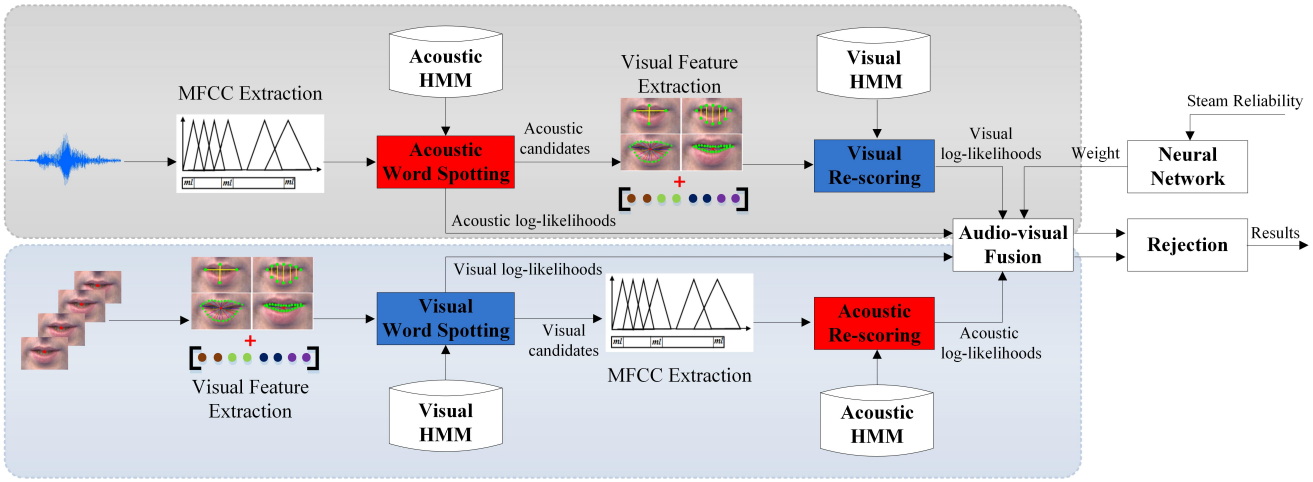


Fig. 2: Audio-visual KWS system showing parallel spotting strategy

small task used to evaluate the performance of beamforming, without specific research on AV-KWS [18]. Additionally, there has been a major focus on English-oriented audio-visual isolated word recognition or connected word recognition, but few studies have addressed AV-KWS. Furthermore, fairly little attention has been paid to AV-KWS for Mandarin, which is one of world's major languages.

In this paper, a decision fusion based AV-KWS strategy is proposed for Chinese Mandarin using a novel lip descriptor. Fig. 2 shows the overall diagram of our AV-KWS system. This paper is a refined and expanded version of our conference proceedings paper [19]. The main contribution of this paper is as follows: 1) For visual speech recognition, a novel lip descriptor is proposed that represents robust and discriminative shape and texture information aiming at suppressing large intra-class variance. Specifically, the state-of-the-art method of facial landmark localization is applied and a set of geometry-based features is proposed based on the localized landmarks. Also a spatiotemporal lip feature is proposed that represents texture changes concerning similarities among them. 2) To adaptively deal with diverse noise conditions and complementarily combine audio speech and visual speech, a parallel two-step KWS strategy based on decision fusion is included. Besides, weights generated using a neural network combine acoustic and visual contributions.

The rest of this paper is organized as follows. In Section II, a novel lip descriptor is presented which consists of the shape difference feature (SDF) and the spatiotemporal lip feature (STLF). In addition, facial landmark localization used for lip region cropping is introduced. Section III presents our adaptive audio-visual integration strategy based on decision-level fusion. Issues in generating integrating weights using a neural network based on stream reliability are also discussed. Then, a parallel two-step keyword spotting strategy as well as an additional step that deals with the time-overlapping situation is described. Experimental results and discussions are provided in Section IV. Finally, conclusions are drawn in Section V.

II. Visual Feature Extraction

Visual speech recognition (VSR), also known as lipreading, is a task of recognizing utterances by analyzing visual recordings of a speaker's talking mouth without any acoustic input [20]. For both AV-ASR and VSR, visual feature extraction is a key research topic and has drawn wide research attention. A review of recent advances in visual speech recognition can be found in [21]. Ziheng Zhou *et al.* proposed a generative latent variable model to provide a compact representation of visual speech data and obtained promising results [20]. In earlier work [22], a practical lipreading system was developed using a simple deterministic model with a low-dimensional manifold, through which visual features extracted from frames of a video could be projected onto a continuous deterministic curve embedded in a path graph. Based on the proposed method, speech videos can be normalized to a standard length. In [23], a spatiotemporal descriptor based on local binary patterns was used for describing isolated phrase sequences, which was originally proposed for texture recognition. Yuru Pei *et al.* presented a random forest manifold technique and applied it to lipreading in color and depth videos [24], in which multiple conventional features like local binary pattern (LBP) and histogram of gradients (HOG) were employed. Generally, most work utilizes either geometry-based or appearance-based visual features directly from other recognition tasks like texture, face, and expression recognition without considering the characteristic of talking mouths. Moreover, most work overlooks the state-of-the-art approaches of facial landmark localization, which can be used to accurately crop the region of the talking mouth.

In the following subsections, first, an advanced facial landmark localization method is employed to promote the process of visual feature extraction, which is crucially important for visual feature extraction in VSR. Second, shape difference features are presented to represent geometric information of lips. Finally, a spatiotemporal lip feature is introduced to capture textures and dynamics of lip movements concerning the following two aspects in the process of speaking: 1) individual variables containing personal identity information leading to large intra-class variance, which are irrelevant to

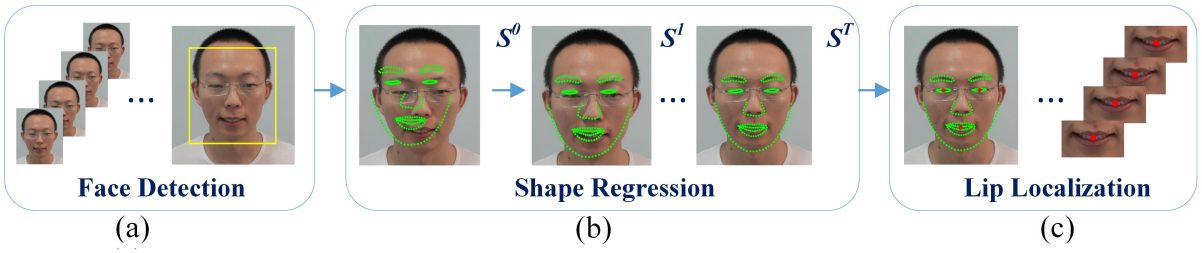


Fig. 3: Flowchart of lip region cropping (a): input utterance video, then perform the coarse face detection; (b): fast shape regression in a coarse to fine manner; (c): face alignment and cropped lip regions.

VSR and need to be suppressed; 2) utterance variables including texture, shape and dynamic changes of the mouth region during speaking, which are the key resource to distinguish different utterances.

A. Lip Region Cropping

Facial landmark localization locates fiducial points on a face image, which is essential for tasks like face recognition and face animation. As a classical method of facial landmark localization, the active appearance model (AAM) is customarily used for the conventional visual front-end in AV-ASR. It is an optimization-based method relying on a parametric model that minimizes parametric errors in the training process. This method is indirect and sub-optimal because smaller parameter errors are not necessarily equivalent to smaller alignment errors. Also, it is well known that AAM is highly sensitive to initialization due to gradient descent optimization. In recent related work [25]–[28], considerable improvements have been made. In particular, a novel regression-based approach without using any parametric shape models is presented in [26]. It shows extraordinary performance in both accuracy and efficiency on three canonical databases BioID [29], LFPW [30] and LFW87 [31]. Based on this approach, a face shape or semantic facial landmark method is employed here for lip region cropping instead of the conventional AAM.

Assume that a face shape $S = [x_1, y_1, ..., x_N, y_N]^T$ consists of N facial landmarks. Given a facial image, the goal of face landmark detection is to estimate a shape S that is as approximate as possible to the true shape $\hat{S}$, *i.e.*, minimizing $\|S - \hat{S}\|$. The basic shape regression framework is as follows. First, boosted regression [32] is used to combine T weak regressors $(R^1, ..., R^t, ..., R^T)$ in an additive fashion. Given a facial image I and initial face shape S^0, each regressor computes a shape increment ΔS from image features and then updates the face shape, which can be formulated as:

$$S^t = S^{t-1} + R^t(I, S^{t-1}), \quad t = 1, ..., T. \tag{1}$$

Given N training examples $\left\{(I_i, \hat{S}_i)\right\}_{i=1}^{N}$, the regressors are sequentially learned until a training error no longer decreases. Each regressor R^t is learned as follows:

$$R^t = \arg\min_{R} \sum_{i=1}^{N} \left\| \hat{S}_i - (S_i^{t-1} + R(I_i, S_i^{t-1})) \right\|, \tag{2}$$

where S_i^{t-1} is the estimated shape in previous stage. Compared with [26], a smart restart approach [27] is added to predict previous failure cases. Referring to [26], [27], the two-level cascaded regression and correlation-based feature selection are adopted in this paper. In regression preprocessing, a rough face box is detected, then the landmark is estimated in a coarse-to-fine manner as illustrated in Fig. 3 (b). Sequentially, geometric centers of eyes and mouths can be detected. As a result, the lip region is cropped depending on the mouth center after normalizing the face using pre-defined ratio parameters for the whole sequence as shown in Fig. 3 (c).

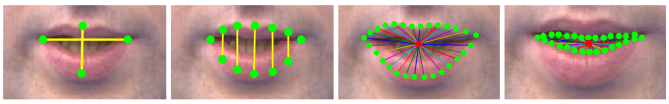


Fig. 4: Four types of shape representation: (a) the lip width and height; (b) shape information in the vertical direction; (c) the outer lip contour; (d) the inner lip contour

B. Shape Difference Feature

Since lip landmarks can be accurately detected due to the efficiency of the shape regression model, a geometry-based feature is concerned to precisely represent the shape, such as lip width, height and contour. The feature named shape difference feature (SDF) is proposed to take full advantage of the derived landmarks.

Given M lip landmarks, four types of representations are developed to comprehensively describe the lip shape here, as shown in Fig. 4 by calculating the Euclidean distance between two landmarks: (a) The lip width and height form a vector denoted as $\boldsymbol{d_1}$; (b) All the vertical distances between corresponding landmarks form a vector $\boldsymbol{d_2}$; (c) The outer lip contour is represented by vector $\boldsymbol{d_3}$ including distances between outer circle landmarks and mouth center; (d) The inner lip contour is represented by vector $\boldsymbol{d_4}$ including distances between inner circle landmarks and mouth center. Denote $\boldsymbol{d}^t = [\boldsymbol{d_1}^T, \boldsymbol{d_2}^T, \boldsymbol{d_3}^T, \boldsymbol{d_4}^T]^T$ as the feature vector from the t-th frame. Concerning the interference from difference of individual mouth appearances, the final shape difference feature vector $\boldsymbol{d}$ is computed as:

$$\begin{aligned} \boldsymbol{d} &= [(\Delta \boldsymbol{d}^1)^T, (\Delta \boldsymbol{d}^2)^T, ..., (\Delta \boldsymbol{d}^T)^T]^T, \\ \Delta \boldsymbol{d}^t &= |\boldsymbol{d}^{t+1} - \boldsymbol{d}^t|, t = 1, ..., T, \end{aligned} \tag{3}$$

where T is the number of frames of the utterance video and $|\cdot|$ means taking the absolute value of each element of $d^{t+1} - d^t$.

C. Spatiotemporal Lip Feature

The mouth appearances of different speakers uttering the same word are diverse, which leads to large intra-class

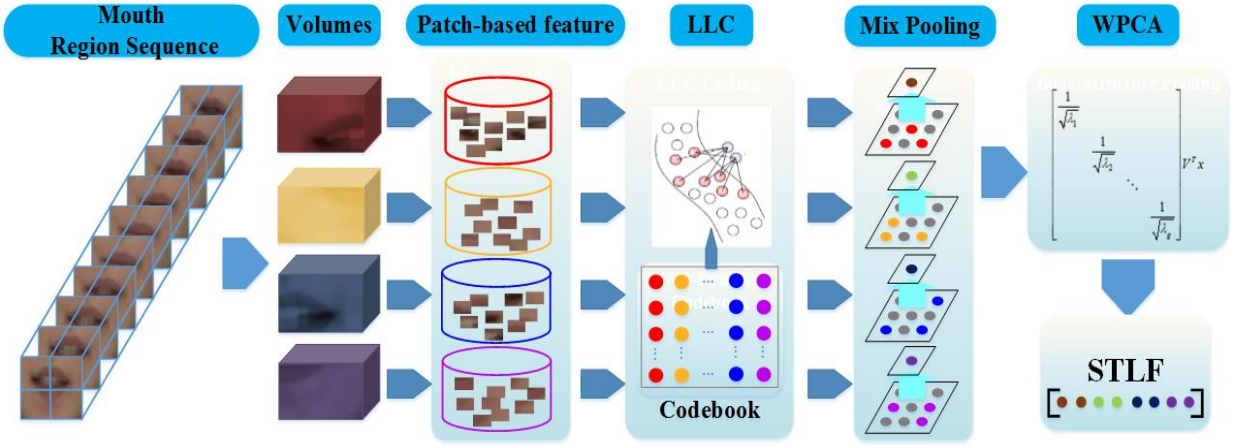


Fig. 5: Process of obtaining spatiotemporal lip feature

variance. To obtain a robust representation of texture and dynamic changes in the detected lip region and to narrow down intra-class distance, a spatiotemporal lip feature (STLF) is proposed with the following steps: 1) parting each cropped mouth region into K blocks and forming into K volumes by putting corresponding blocks together in an utterance video; 2) extracting low-level features from sampling patches; 3) employing locality-constrained linear coding (LLC) [33] to encode low-level features into higher ones and implementing the proposed mix pooling; 4) employing whitened principle component analysis (WPCA) to narrow down intra-class variations. The overall procedure is illustrated in Fig. 5.

As an aid to illustration, we introduce the term *texton* here which is defined as a mini-template that consists of a varying number of image bases with some geometric and photometric configurations [34]. To enhance textons and suppress Gaussian noise, a difference of Gaussians (DoG) filter is first applied to each cropped mouth region. Sequentially, the low-level patch-based features $\boldsymbol{p}_i \in \mathbb{R}^D, i \in \{1, 2, ..., N\}$ are extracted from the volume, where D is the dimension of the feature vectors obtained.

Locality-constrained Linear Coding (LLC): Although the low-level patch-based features are able to capture subtle textons in the mouth region, it is not optimal to use them directly without coding since great similarities among textons may result in low discriminability. To derive a more robust and discriminative descriptor, LLC [33], a fast and effective coding strategy is used to encode the low-level patch-based features, which shows better performance than common coding schemes such as vector quantization and sparse coding.

When it comes to coding, a codebook needs to be employed and K-Means [35] is utilized. Denote the over-complete sub-codebook as $\boldsymbol{B_k}$, which is learned from the low-level patch-based features in the corresponding volume k. Therefore, the codebook $\boldsymbol{B}$ is constructed as follows:

$$\begin{aligned} \boldsymbol{B} &= \{\boldsymbol{B}_k | k = 1, ..., K\}, \\ \boldsymbol{B}_k &= [\boldsymbol{b}_{k,1}, \boldsymbol{b}_{k,2}, ..., \boldsymbol{b}_{k,M}] \in \mathbb{R}^{D \times M}, \end{aligned} \tag{4}$$

where M is the number of entries in the codebook and $M \gg D$. Then, the low-level patch-based features can be encoded using the following criteria, which has an analytical solution:

$$\begin{aligned} \min_{\boldsymbol{C}} \sum_{i=1}^{N} \|\boldsymbol{p}_i - \boldsymbol{B}_k \boldsymbol{c}_i\|^2 + \lambda \|\boldsymbol{d}_i \odot \boldsymbol{c}_i\|^2, \\ s.t. \quad \mathbf{1}^T \boldsymbol{c}_i = 1, \forall i, \end{aligned} \tag{5}$$

where $\odot$ denotes element-wise multiplication, $\mathbf{c}_i$ is the reconstructed vector, i.e., the code, $\boldsymbol{C} = [\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_N]$ the set of codes and $\boldsymbol{d}_i \in \mathbb{R}^M$ is a locality adaptor which assigns different proportion for each basis according to similarity defined as:

$$\begin{aligned} \boldsymbol{d}_i &= \exp\left(\frac{dist(\boldsymbol{p}_i, \boldsymbol{B}_k)}{\sigma}\right), \\ dist(\boldsymbol{p}_i, \boldsymbol{B}_k) &= [dist(\boldsymbol{p}_i, \boldsymbol{b}_{k,1}), ..., dist(\boldsymbol{p}_i, \boldsymbol{b}_{k,M})], \end{aligned} \tag{6}$$

where $dist(\boldsymbol{p}_i, \boldsymbol{b}_{k,j})$ is the Euclidean distance between $\boldsymbol{p}_i$ and $\boldsymbol{b}_{k,j}$ and σ is used to adjust the weight decay speed of the locality adaptor. In (5), the item $\sum_{i=1}^{N} \|\boldsymbol{p}_i - \boldsymbol{B}_k \boldsymbol{c}_i\|^2$ representing reconstruction error is minimized and the regularization term is used to derive more discriminative reconstruction, generating similar codes for similar textons. The analytical solution of (5) is as follows:

$$\widetilde{\boldsymbol{c}}_i = (\boldsymbol{C}_i + \lambda diag(\boldsymbol{d})) \backslash \mathbf{1}, \tag{7}$$

$$\boldsymbol{c}_i = \widetilde{\boldsymbol{c}}_i / \mathbf{1}^T \widetilde{\boldsymbol{c}}_i, \tag{8}$$

where $\boldsymbol{C}_i = (\boldsymbol{B}_k - \mathbf{1}\boldsymbol{p}_i^T)(\boldsymbol{B}_k - \mathbf{1}\boldsymbol{p}_i^T)^T$ denotes the data covariance matrix.

Mix Pooling: After deriving a set of high dimensional codes, feature pooling is applied here to obtain statistical information in each volume. This also improves the robustness and makes subsequent calculations more orderly. Generally, there are two common pooling strategies, namely sum pooling and max pooling. Specially, the feature of the k-th volumes can be obtained as:

$$sumpooling : \boldsymbol{x}_k = \sum_{t=1}^{T} \sum_{\boldsymbol{c}_i \in S_k^t} \boldsymbol{c}_i, \tag{9}$$

$$maxpooling : \boldsymbol{x}_k = \max_{t \in [1,...,T]} \max_{\boldsymbol{c}_i \in S_k^t} \boldsymbol{c}_i, \tag{10}$$

where T is the number of frames of the utterance video, S_k^t is the t-th frame from the k-th volume, and $\mathbf{c}_i$ represents the i-th code in S_k^t.

Max pooling features can capture the salient properties of a region [36], thus it can be employed in each divided block. However, max pooling features among blocks in a volume could result in losing dynamic information of the talking process. Thus mix pooling is proposed to maintain the dynamic information between frames in a volume while obtaining salient textons in a block, which is defined as follows:

$$mixpooling : \boldsymbol{x}_k = \sum_{t=1}^{T} \max_{\boldsymbol{c}_i \in S_k^t} \boldsymbol{c}_i. \tag{11}$$

WPCA: As the feature vector $\boldsymbol{x}_k$ derived in the above step is of high dimension, a compact representation needs to be explored. Commonly, principle component analysis (PCA) is used to the reduce feature dimension by only preserving the eigenvectors corresponding to large eigenvalues. Considering that there are great differences among utterances of the same keyword by different people, PCA may have a tendency to magnify the difference. To suppress the difference from personal variations, whitened PCA (WPCA) is applied through the following steps: 1) Map the feature $\boldsymbol{x}_k$ to intra-class subspace by calculating the intra-class covariance matrix $\boldsymbol{C}_k \in \mathbb{R}^{M \times M}$ of k-th volumes of all training examples. 2) Let $\Lambda = \{\lambda_1, ..., \lambda_g\}$ be the g largest eigenvalues and $\boldsymbol{V} = [\boldsymbol{v}_1, ..., \boldsymbol{v}_g]$ be the corresponding eigenvectors. 3) Obtain a compact representation $\boldsymbol{y}_k$ for the k-th volume of an utterance video as follows:

$$\boldsymbol{y}_k = diag(\lambda_1^{-1/2}, ..., \lambda_g^{-1/2}) \boldsymbol{V}^T \boldsymbol{x}_k. \tag{12}$$

Note that the features are multiplied by the inverse of the eigenvalues, which suppresses the responses from larger eigenvalues. Therefore, the difference from individual variables, *i.e.*, intra-class dissimilarity is reduced. Thus the feature vector for all K volumes of an utterance can be represented as:

$$\boldsymbol{y} = [\boldsymbol{y}_1{}^T, \boldsymbol{y}_2{}^T, ..., \boldsymbol{y}_k{}^T]^T. \tag{13}$$

We also apply WPCA to SDF which was introduced in the previous subsection, and it remains the same notation $\boldsymbol{d}$. So the combination of two features can be represented as:

$$\boldsymbol{f} = [\boldsymbol{d}^T, \mu \cdot \boldsymbol{y}^T]^T, \tag{14}$$

where μ is an adjustment factor to balance the relative importance of two features. To select a proper μ, a method similar to that in [37] is adopted.

III. Parallel Two-step KWS Strategy Based on Decision Fusion

A. Adaptive Audio-visual Integration

For AV-KWS, the audio-visual integration module plays an important role. Obviously, contributions of acoustic information and visual information are different under various noisy conditions. Therefore, the decision on how to fuse acoustic and visual information significantly influences the final performance. Generally, there are two broad fusion categories: feature-level fusion and decision-level fusion [14]. Feature-level fusion directly concatenates the features of the two modalities into a larger feature vector in a plain way, or it adopts some appropriate transformations. Recognition is then conducted using a single classifier based on the concatenated feature vector as shown in Fig. 6 (a). Alternatively, in decision-level fusion, audio and video modalities are integrated at the classifier output level as shown in Fig. 6 (b). Specifically, decision fusion is classified into three possible categories, namely "early" integration (multi-stream HMM), "intermediate" integration and "late" integration integration. These three categories respectively correspond to classification at state-level, word-level and utterance-level [11].

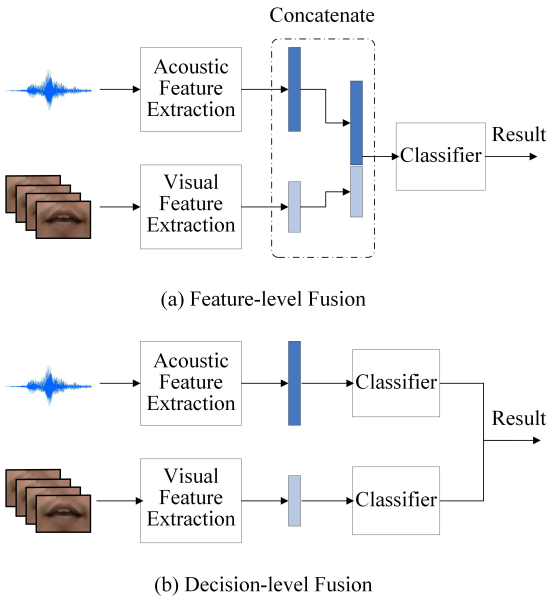


Fig. 6: Framework of feature-level fusion and decision-level fusion

Compared with feature-level fusion, decision-level fusion approaches have important advantages in handling different noisy conditions [11], [14]. Feature-level fusion concatenates acoustic features and visual features into a larger feature vector with higher dimensionality, thus more training data are needed to ensure adequate probabilistic modeling. In contrast to feature-level fusion, decision-level fusion can explicitly model the reliability of two modalities, which is of great significance since the discrimination power of the two modalities may vary widely. According to differing noisy conditions, integrating weights are relatively easy to generate. This facilitates control of the contributions of the two modalities using decision-level fusion, independently handling the two modalities.

In this paper, late integration of decision fusion is employed in order to cope with different noise conditions and develop a noise-robust AV-KWS system. Also, conventional AV-KWS based on HMM is utilized, where acoustic HMMs and visual HMMs are respectively trained to provide corresponding modality likelihoods of a given multi-media source. Integrated scores can be obtained by linearly combining acoustic and visual log-likelihoods of keyword candidates using the appropriate weights as follows [38]:

$$\log p(O_{AV}|\lambda_i) = \gamma \log p(O_A|\lambda_i^A) + (1-\gamma) \log p(O_V|\lambda_i^V), \tag{15}$$

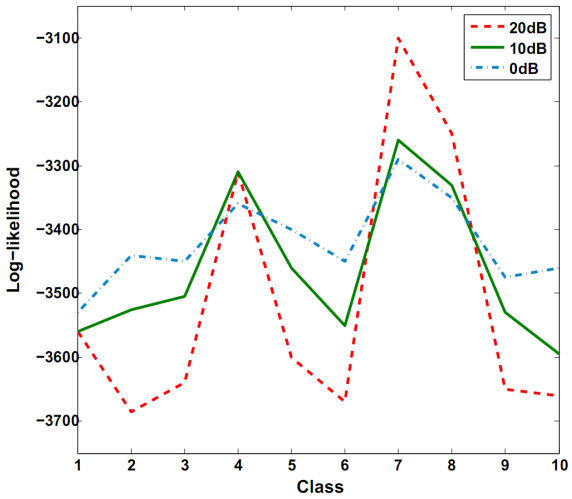


Fig. 7: Log-likelihoods of HMMs under different noisy conditions

where γ denotes the integration weight with a value between 0 and 1. O_A and O_V are the acoustic and visual feature sequences of a keyword candidate while λ_i^A and λ_i^V are the acoustic and visual HMM of keyword i. Items $\log p(O_A|\lambda_i^A)$ and $\log p(O_V|\lambda_i^V)$ represent the corresponding acoustic and visual log-likelihood.

For a specific environment, an integration weight with a constant value can be estimated according to optimal performance under specified conditions. However, when audio-visual environments change dramatically, a fixed integration weight is insufficient to cope with the noise-varying conditions. Confronted with diverse noisy conditions, a crucial issue lies in obtaining adaptive integration weights based on the reliabilities of two modalities [39]. Fig. 7 shows a typical example of the output log-likelihoods given a speech with different noise conditions. It can be observed that the output log-likelihoods of each HMM in quiet environments display great differences while small differences are noted when environments are noisy. A large difference reflects less ambiguity and larger certainty for recognition. Since the output log-likelihood of HMMs can reflect current noisy conditions, it has been commonly used as a reliability measure [40]. Various forms of reliability measures based on log-likelihoods have been implemented by researchers [41]–[43]. In this paper, the average difference against the best hypothesis with the maximum log-likelihood [42] is adopted as the reliability measure:

$$D = \frac{1}{N-1}\sum_{i=1}^{N}\left(\max_j L_j - L_i\right), \tag{16}$$

where $L_i = \log p(O|\lambda^i)$ is the output log-likelihood of the i-th HMM and N denotes the number of HMMs.

Studies show that the average difference against the maximum log-likelihood (*Diffmax*) in (16) has the best recognition performance under different noisy conditions from an overall point of view [14]. Lewis and Powers pointed out that the intrinsic errors of other dispersion forms in measuring reliabilities leading to their inferiority to Diffmax [44]. Therefore, the reliability measure of average difference against the maximum log-likelihood is used in this paper.

A neural network is then utilized to map the two input reliabilities to the optimal weight. Regarding a keyword candidate with a starting and ending time, corresponding reliabilities of each modality (D_A and D_V) can be effectually obtained. Integrating weight γ can be calculated by the function f modeled by the neural network for a given pair of acoustic, visual reliabilities (D_A, D_V) as follows:

$$\gamma = f(D_A, D_V). \tag{17}$$

In order to obtain adaptive weights for various conditions, acoustic speech utterances with different SNRs and visual speech utterances with different image resolutions are utilized to train the neural network. The trained neural network can generate the optimal weight of a keyword candidate for different conditions, not limited to the conditions used for training.

The precise training process proceeds as follows: (1) Calculate D_A and D_V of a given labeled keyword (The keyword in the utterance is artificially labeled). (2) Exhaustively search the optimal weight over the space of [0,1] with a step of 0.01 and check whether the recognition result using the particular weight value is correct. (3) Train the neural network using the input reliabilities and the corresponding optimal weights.

B. Two-step Keyword Spotting Strategy

To determine the benefit of visual modality to KWS using the adaptive weights, the conventional HMM-filler based KWS is employed. This method is primarily used in application fields such as dialogue systems, and command control and information consultation. More specifically, our KWS system applies the conventional two-stage strategy: picking out possible keyword candidates to include true keywords embedded in unconstrained speech in the first stage, and rejecting false alarms in the second stage. Since acoustic recognition performance drops significantly in acoustically noise conditions, the strategy of merely performing visual re-scoring on the acoustic candidate is abandoned. Alternatively, a parallel two-step recognition is introduced to complementarily make full use of two modalities, as shown in Fig. 2.

Step 1: With the trained acoustic and visual keyword HMMs as well as the filler models, acoustic and visual keyword searching are first conducted in parallel on the tested speech, generating a number of acoustic keyword candidates as well as visual keyword candidates with the corresponding log-likelihoods.

Step 2: For a keyword candidate obtained by either modality with a starting and ending time, re-scoring based on the other modality of the keyword is then performed since acoustic keyword candidates may not be the same as the visual ones, especially when acoustic environments are too noisy. Therefore, each candidate receives an acoustic and a visual log-likelihood.

With the corresponding acoustic reliability D_A and visual reliability D_V being calculated, the trained neural network takes D_A, D_V as input factors and outputs the optimal weight. Next, integrated scores of keyword candidates can be obtained by linearly combining the acoustic and visual log-likelihoods

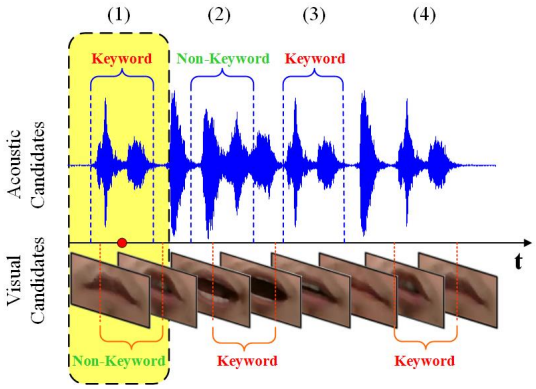


Fig. 8: Additional step to deal with the overlapping acoustic and visual candidates

using the estimated weights in (15). Finally, rejection is implemented based on the integrated scores of each keyword candidate to remove false alarms. Taking the background noise into consideration, rejection based on likelihood ratio [45] (or log likelihood difference) denoted as L_r is utilized due to its robustness to noise, as follows:

$$L_r = \log p(O_{AV}|\lambda_i) - \log p(O_{AV}|Filler), \qquad (18)$$

where $\log p(O_{AV}|\lambda_i)$ is the integrated log likelihood of the keyword model λ_i and $\log p(O_{AV}|Filler)$ is the integrated log likelihood of the filler model. Similarly, $\log p(O_{AV}|Filler)$ can be calculated based on the filler model by linearly combining corresponding acoustic and visual log likelihoods. The candidate is accepted as a true keyword when its log likelihood ratio is greater than a threshold, otherwise it is considered as a false alarm and is rejected.

Conventionally, recognition result analysis is performed by comparing it with an artificially labeled reference after keyword verification. As depicted in Fig. 8, some acoustic and visual keyword candidates are directly removed as false alarms in the rejection step (case (2) in Fig. 8). For those remaining candidates after rejection, an additional step should be taken since acoustic and visual keyword candidates may overlap in time. Therefore, a criterion is required to deal with the situation. For each acoustic and visual keyword candidate with a corresponding time region and integrated loglikelihood, if the middle time point of one modality keyword candidate falls within the time region of the other modality keyword candidate, it is regarded as an overlapping instance. Therefore, the candidate with greater integrated loglikelihood is determined to be the true keyword while the other is regarded as a false alarm (case (1) in Fig. 8). For other cases (acoustic and visual keyword candidates do not overlap in time), candidates are directly determined to be true keywords (case (3), (4) in Fig. 8).

IV. Experiments and Discussions

A. Visual Speech Recognition

This subsection describes experiments that are implemented on a visual-only benchmark database OuluVS [23] to validate the proposed visual features. OuluVS consists of 20 subjects uttering 10 phrases five times with resolution of 720×576 pixels. The phrases are listed in Table I. Visual speech recognition here in particular is to classify the entire phrase. For preprocessing, the lip regression model is applied and a 100×80 lip region is cropped off from each video frame. Due to the difference of subjects' speeds of utterance, the same path graph based video normalization scheme in [22] is employed. Concretely, all the utterances are normalized to be 30 frames long. Experiments are carried out in the speaker-independent way, that is the training and testing data are from different subjects. Leave-one-subject-out is employed by training on $N-1$ (N is the total number of subjects in the database) speakers, while testing on the remaining one. Moreover, an SVM classifier is trained for each pair of phrases, then the majority voting scheme is adopted to decide which phrase the test utterance video belongs to.

TABLE I: Phrases in OuluVS dataset

C1	"Excuse me"	C6	"See you"
C2	"Goodbye"	C7	"I am sorry"
C3	"Hello"	C8	"Thank you"
C4	"How are you"	C9	"Have a good time"
C5	"Nice to meet you"	C10	"You are welcome"

Experiment 1: To evaluate STLF, several parameters first need to be discussed. The DoG filter is set to $\sigma = 2$ while the sampling patch size is set to 6×6 pixels. The dimension of WPCA is set to $g = 60$ while PCA with the same dimension is also tested for an alternative way. To obtain a favorable volume size, four segmentation ways, $K = 2 \times 2$, 4×2, 5×2 and 8×4 are considered. Also, different codebook sizes $M = 64, 128, 512$ and 1024 are evaluated. In addition, to verify the proposed mix pooling strategy, max and sum pooling are also implemented. As shown in Fig. 9 (a), the most advantageous segmentation way is 8×4 and mix pooling achieves more favorable performance than the other two pooling strategies. Mix pooling with 8×4 segmentation achieves the highest recognition accuracy of 76.25%. This is consistent with the previous hypothesis that the mix pooling selects discriminative features in each frame through max pooling while preserving the changes between frames through sum pooling. From Fig. 9 (b), it can be seen that WPCA has a more competitive performance than PCA, since it has the ability of reducing intra-class variation. The codebook size is set to 512 in the following discussions for a good tradeoff between the performance and the efficiency.

Experiment 2: To explore the performances of SDF, STLF and their combination, they are compared with state-of-the-art lipreading methods proposed in [22]–[24]. Further, discrete cosine transform (DCT), a conventional feature extraction method in VSR, is also employed as an elementary baseline. For STLF, the optimal parameters are adopted, that is using 8×4 segmentation, mix pooling and WPCA. In Table II and Fig. 10, "SDF + STLF" represents concatenation of the two feature vectors, where the adjustment factor μ is set to 1.2 using coarse to fine procedure. For DCT, 13 most important coefficients are utilized and their first and second

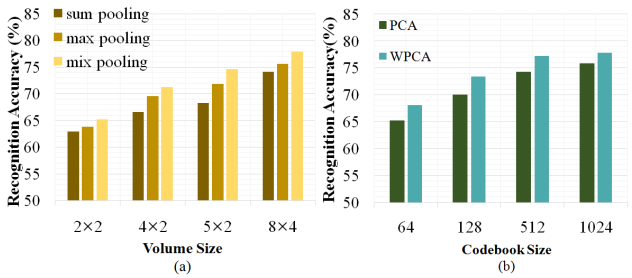


Fig. 9: (a) Comparisons of STLF on OuluVS with different segmentations and pooling strategies with codebook size $M = 512$ using WPCA; (b) Comparisons of STLF with different codebook sizes and dimension reduction methods in segmentation 8×4 using mixpooling.

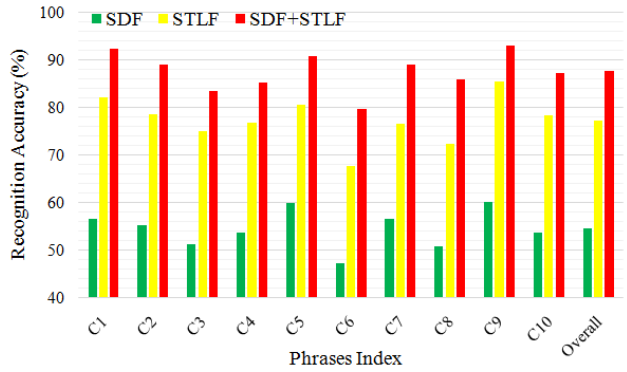


Fig. 10: Phrases recognition comparison of different features in the speaker-independent way on OuluVS database

derivatives are included. More details of the employed DCT can be found in [46]. Experimental results in Fig. 10 show that the appearance-based feature STLF demonstrates a far more competitive performance than the geometric-based feature SDF. This is due to the fact that STLF extracts from the pattern level intrinsically, where LLC retains the ability to generate similar codes for similar textons, meanwhile mix pooling preserves most salient features and WPCA reduces the intra-class variance. Despite the accurate detection of lip landmarks, SDF still largely contains interference due to differing magnitudes of mouth opening, and differing speeds and accelerations of mouth movements. However, we regard SDF as a by-product of facial landmark localization. It is not computationally expensive but can be used profitably as a complementary feature. As Fig. 10 shows, label "C1" to "C10" indicate different short sentences in Table I, where "C6" denotes "See you" getting the poorest performance. This may be caused by some quick pronouncing of "See" in the database as no adequate feature is captured. In Table II, the combination of SDF and STLF outperforms the methods proposed in [22], [23] and nearly matches [24] while the performance of DCT is far behind. Comparatively, Pei *et al.* [24] employed multi-modal data HOG, LBP and trajectories shape as well as depth information. Additionally, the computational complexity of STLF is $O(M + N)$, which is linearly related to the size of the codebook and the number of sampled patches.

B. Audio-visual Keyword Spotting on PKU-AV

As only a few common databases are available for AV-ASR [11], [21], [47] and the existing audio-visual databases are

TABLE II: Lipreading performances on OuluVS

Datasets	Methods	Accuracy (%)
OuluVS	DCT	37.09
	SDF	54.35
	STLF	76.25
	SDF+STLF	**87.55**
	Zhao *et al.* [23]	58.85
	Zhou *et al.* [22]	81.30
	Pei *et al.* [24]	89.70

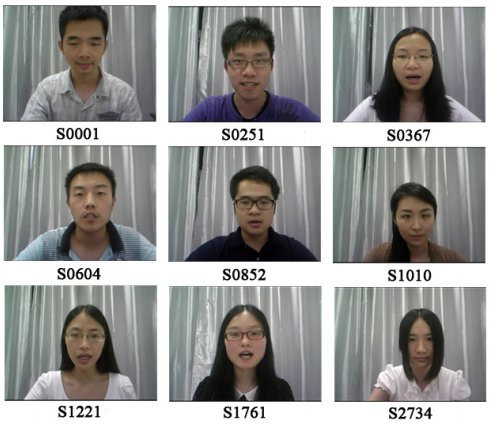


Fig. 11: Exemplar video frames in PKU-AV

rarely concerned with AV-KWS of Mandarin, a novel audio-visual database named PKU-AV is established to conduct experiments considering AV-KWS of Mandarin.

This audio-visual database contains 20 subjects (12 male Asians and 8 female Asians) and there are 300 utterances for each subject. Concerning the integrated functions of the HRI including such tasks as smile detection, gender recognition, and age estimation recognition, 30 frequently used keywords are defined. The 30 Mandarin keywords translated into English are as follows: "Forward", "Backward", "Left", "Right", "Turn around", "Fast", "Slow", "Start", "Stop", "Continue", "Pay attention", "Help", "Gender", "Age", "Identification", "Smile", "Expression", "Hands up", "Action", "Tracking", "Localization", "Photo", "Display", "Play", "Record", "Inquiry", "Selection", "Function", "Abstract", "Update". One example of a full utterance including the keyword "Localization" is: "*Please carry out sound source localization*". In accordance with the pronunciation characteristics of Chinese Mandarin, for this KWS task, six consonantal and five vocalic visemes are considered from 47 possible Chinese phonemes. Our database is constructed with the addition of artificial acoustic noise. First, an original database is collected in an acoustically quiet environment under controlled normal light conditions. Video images are collected at 20 frames per second with a resolution of 640×480 under the restriction that the mouth region is not occluded. Audio speech is synchronously recorded at the sampling rate of 16 kHz and 16 bits per sample. Fig. 11 depicts some representation video frames in the database.

To allow speaker-independent recognition, the database PKU-AV is divided into three sets: (1) Training sets composed of original AV data from 7 subjects are used to train acoustic and visual HMMs. (2) Held-out sets consisting of AV data from 6 subjects are utilized to train the neural network. The noisy AV data has acoustic SNRs of 20dB, 10dB and 0dB by

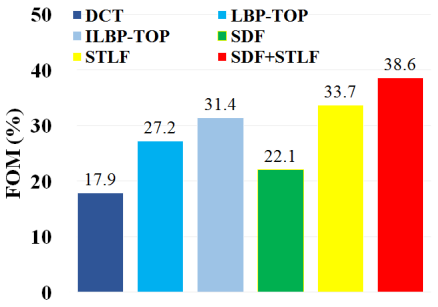


Fig. 12: FOM performances of different visual features for keyword spotting

addition of white noise. (3) Test sets composed of noisy AV data from 7 subjects are used to evaluate performances of our AV-KWS system under various conditions, including different acoustic noise (white and babble noise) with different SNRs (20dB, 15dB, 10dB, 5dB and 0dB) as well as various image resolutions ("100×80", "50×40" and "25×20").

For all of the following experiments, the HMM-filler based KWS method in [7] is employed and the commonly used Mel-frequency cepstral coefficients (MFCCs) and its delta as well as delta delta are extracted for acoustic features using the HTK toolbox [48]. Extracted acoustic and visual features are individually used to train corresponding HMM classifiers. To illustrate, both acoustic and visual keyword HMMs are trained based on the whole word since keyword dependence can improve the performance [49]. Sub-word units are used as corresponding filler models and each sub-word unit is modeled with a 3-state HMM with each state containing eight Gaussian components. Experimentally, the number of hidden neurons of the neural network is set to six and figure of merit (FOM) is utilized for measurement. FOM is defined as the average percentage of correctly detected keywords as the threshold is varied from one to 10 false alarms per keyword per hour [7], [50]. For visual preprocessing, all the faces in PKU-AV are successfully detected and the lip regression model is applied to obtain a 100×80 mouth region from each video frame.

Experiment 3: Performances of SDF, STLF and their combination on PKU-AV are tested for keyword spotting through HMM. For DCT, STLF and its combination with SDF, the parameter settings are the same as in Experiment 2. Comparisons with previous methods [19] are also carried out with the same experimental protocols to explore the effectiveness of our proposed visual features. As shown in Fig. 12, both SDF and STLF outperform DCT while the improved performance of STLF is almost the double that of DCT. The combination of SDF and STLF achieves the optimal result followed by STLF and ILBP-TOP proposed in [19]. Moreover, it can be observed that the performance of individual STLF outperforms that of LBP-TOP proposed in [23] and ILBP-TOP, which demonstrates that STLF is a more compact and discriminative representation. After combing STLF with SDF, the performance is improved by about 5% as SDF contains geometric supplementary information to STLF.

Experiment 4: In this experiment, the visual part and the audio part are integrated based on the decision level. Experiments are carried out to explore performances of the audio-only, vision-only and audio-visual KWS using SDF, STLF and their combination under various acoustic noise conditions (white noise and babble noise). Fig. 13 indicates that the performance of audio-only recognition significantly degrades as speech becomes noisier. Vision-only performance appears the same for all the SNR conditions, which can be explained by the invariance of visual conditions. In addition, the integration of acoustic and visual modality significantly improves the noise robustness of the KWS system. Clean speech utterances corrupted by white noise with SNR of 20dB, 10dB and 0dB are used to train the neural network. And tests conducted on white noise and babble noise speech utterances at various SNR conditions (20dB, 15dB, 10dB, 5dB and 0dB) show that this approach also works well for untrained noise conditions including different noise levels as well as noise types. Moreover, audio-visual KWS using STDF demonstrates more robustness performance than using SDF when SNR declines, while using their combination obtains the most favorable result.

TABLE III: Audio-only, vision-only and audio-visual performances in terms of FOM (%) using different fusion methods

SNR(dB)	20	15	10	5	0
Audio-only	74.7	57.4	39.2	18.6	6.4
Vision-only	38.6	38.6	38.6	38.6	38.6
Feature-level AV [17]	77.2	65.9	49.2	40.9	37.5
Decision-level AV	80.5	69.1	58.2	43.7	40.8

Experiment 5: Next, a comparison is made between the AV-KWS performance based on decision-level fusion using adaptive weights (using "SDF+STLF" as visual features) and the feature-based audio-visual keyword spotter proposed in [17] on the database (white noise corrupted acoustic speech and original visual speech). Since very little work has been implemented for audio-visual keyword spotting, this method is compared to the most related one [17]. As shown in Table III, this approach is more robust to noise than that of [17]. The integrated audio-visual performance is at least equal to or better than that of unimodality while the integrated performance in [17] worsens compared to vision-only performance at SNR of 0dB. An explanation of this phenomenon of the feature-level fusion approach is that under extremely low SNR, the audio information introduces harmful cues and may degrade the overall performance of audio-visual fusion. To offset this, the contribution of acoustic and visual modality is combined using adaptive weights according to current SNR conditions in the decision-level fusion method, which may complementarily produce a better overall performance.

Experiment 6: Relatively free movement should be allowed in a friendly face-to-face HRI, which may lead to different face sizes concerning AV-KWS. In order to explore the influence of face size changes on performance, experiments are carried out on videos with different resolutions. Fig. 14 shows the performances of audio-only KWS and AV-KWS on different face size changes using combinations of SDF and STLF, where "100×80", "50×40" and "25×20" denote different resolutions of the mouth region. It can be observed that the

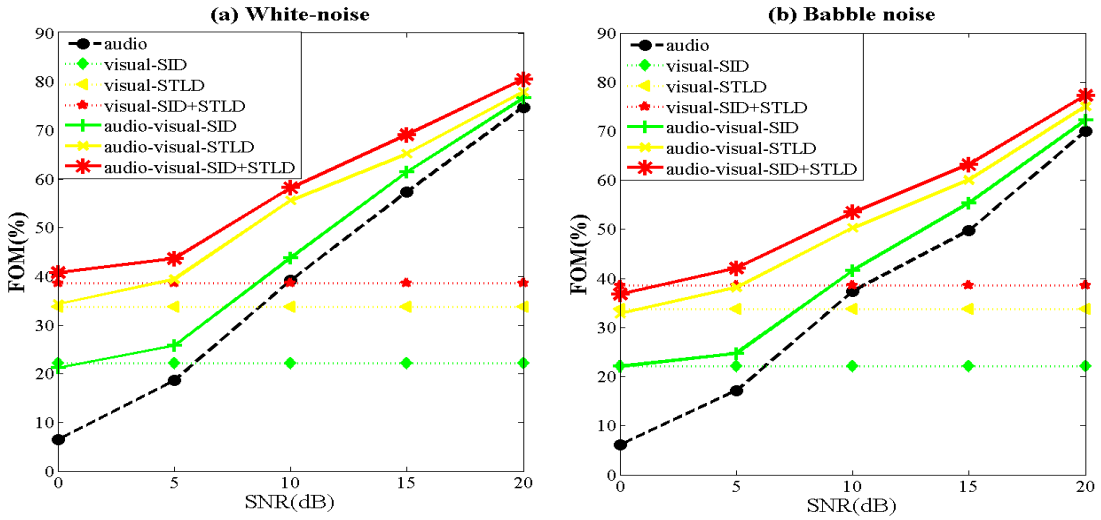


Fig. 13: Recognition performances of the audio-only, vision-only and audio-visual KWS system under white noise and babble noise conditions. (a) performance of white noise condition using SDF, STLF and their combination. (b) performance of babble noise condition using SDF and STLF their combination.

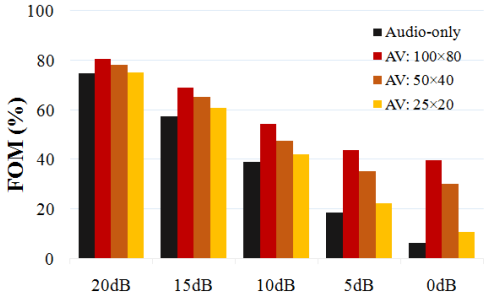


Fig. 14: FOM performances of different image resolutions

performances of AV-KWS degrade along with the decline of image resolution, which can be explained by the loss of texture information as well as motion information. In addition, performances of audio-visual KWS are more competitive than those of traditional audio-only KWS, especially under conditions with low SNR. Even when the image resolution is down-sampled to "25×20", audio-visual integration improves the general performance compared to audio-only KWS. These performances of different image resolutions caused by changes of image size can be utilized to guide users to change positions in a face-to-face HRI for more effective interaction. For instance, when performance drops dramatically due to the low image resolution, a reminder can appear that the current situation is not optimal for interaction and the robot perhaps should move closer to the user.

C. HRI in Real Environments

This AV-KWS system is also attached to a robot platform. Nicknamed Pengpeng II, it is an HRI oriented mobile robot system as depicted in Fig. 15. It is also used as a platform for sound source localization [51]. In the HRI environment, there are various kinds of noise in the hall including air-conditioning noise, motor noise from the robot itself, human voices and so on. In contrast to database experiments, the SNR cannot be strictly controlled due to the complexity of noise.

Experiment 7: In the following experiment, three kinds of noise intensity are estimated: weak noise with an average SNR of 18.6dB, moderate noise with an average SNR of 11.2dB and strong noise with an average SNR of 4.8dB. An EPD module is also necessary for HRI in real environments. Therefore, a GMM-EPD [52] is used for detecting speech activity. Ten lab members participate in the experiment and each speaks 50 sentences including 10 keywords defined in PKU-AV. In order to interact with the robot in a friendly way, the subjects are expected to interact within a range of 1.0 to 1.5 meters since our visual method has the capability to deal with low-resolution images to a certain degree. Average performances are shown in Table IV.

Fig. 15: HRI oriented mobile robot Pengpeng II

Table IV shows that performances of audio-visual keyword spotting based on decision fusion degrade compared to experiments on the database PKU-AV in similar SNRs. The main reason for the drop in performance is the complexity of

TABLE IV: True positive rate (TPR) and false positive rate (FPR) of our audio-visual keyword spotting on Pengpeng II

SNR(dB)	**18.6**	**11.2**	**4.8**
TPR	70.2%	47.7%	32.5%
FPR	7.1%	5.8%	4.9%

both acoustic and visual conditions: various noise, illumination changes, movements of heads, distance changes and so on. Moreover, in strong noise conditions when audio information is unreliable, visual speech recognition significantly improves the performance in real environments.

V. Conclusions

In order to obtain more robust HRI under various noisy conditions, this paper develops an audio-visual keyword spotter using novel visual features. The state-of-the-art facial landmark localization method is utilized to accurately crop and align lip regions. To make full use of the detected landmarks, a geometric feature SDF is designed as a complementary feature. The proposed STLF takes lip texton similarities into account and works to reduce intra-class variances. A parallel two-step recognition, based on both acoustic and visual modality, is also conducted in order to make the best use of the two modalities under various conditions. Experimental results validate the effectiveness of the proposed features as well as their combination. In addition, this audio-visual integration based on decision level improves the noise robustness of the keyword spotter. This audio-visual keyword spotter's ability to deal with untrained noisy conditions including different noise levels as well as noise types is strongly confirmed.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC, no. 61340046, 60875050, 60675025), the National High Technology Research and Development Programme of China (863 Programme, no. 2006AA04Z247), Guangdong Natural Science Foundation of China (Grant 2015A030311034) and the Specialized Research Fund for the Doctoral Programme of Higher Education (Grant 20130001110011).

References

[1] H. Soltau, G. Saon, L. Mangu, H.-K. Kuo, B. Kingsbury, S. Chu, and F. Biadsy, "Automatic speech recognition," in *Natural Language Processing of Semitic Languages*, pp. 409–459, Springer, 2014.

[2] N. Morgan, "Deep and wide: Multiple layers in automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 7–13, 2012.

[3] H. Jiang, "Discriminative training of hmms for automatic speech recognition: A survey," *Computer Speech & Language*, vol. 24, no. 4, pp. 589–608, 2010.

[4] A. A. Abdelhamid, W. H. Abdulla, and B. MacDonald, "Wfst-based large vocabulary continuous speech decoder for service robots," in *International Conference on Imaging and Signal Processing for Healthcare and Technology*, pp. 150–154, 2012.

[5] M. Akbacak, L. Burget, W. Wang, and J. van Hout, "Rich system combination for keyword spotting in noisy and acoustically heterogeneous audio streams," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 8267–8271, IEEE, 2013.

[6] P. Motlicek, F. Valente, and I. Szoke, "Improving acoustic based keyword spotting using lvcsr lattices," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 4413–4416, IEEE, 2012.

[7] I. Szoke, P. Schwarz, P. Matejka, L. Burget, M. Karafiát, M. Fapso, and J. Cernocky, "Comparison of keyword spotting approaches for informal continuous speech," in *Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, 2005.

[8] J. Keshet, D. Grangier, and S. Bengio, "Discriminative keyword spotting," *Speech Communication*, vol. 51, no. 4, pp. 317–329, 2009.

[9] V. Mitra, J. van Hout, H. Franco, D. Vergyri, Y. Lei, M. Graciarena, Y.-C. Tam, and J. Zheng, "Feature fusion for high-accuracy keyword spotting," in *International Conference on Acoustics, Speech and Signal Processing*, pp. 7143–7147, IEEE, 2014.

[10] C. Weng and B.-H. F. Juang, "Discriminative training using non-uniform criteria for keyword spotting on spontaneous speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 2, pp. 300–312, 2015.

[11] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *Issues in visual and audio-visual speech processing*, vol. 22, pp. 356–396, 2004.

[12] S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "Audio-visual fusion and tracking with multilevel iterative decoding: framework and experimental evaluation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 882–894, 2010.

[13] J.-S. Lee and C. H. Park, "Robust audio-visual speech recognition based on late integration," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 767–779, 2008.

[14] J.-S. Lee and C. H. Park, "Adaptive decision fusion for audio-visual speech recognition," *Speech recognition, technologies and applications*, pp. 275–296, 2008.

[15] A. H. Abdelaziz, S. Zeiler, and D. Kolossa, "Learning dynamic stream weights for coupled-hmm-based audio-visual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 5, pp. 863–876, 2015.

[16] V. Estellers, M. Gurban, and J.-P. Thiran, "On dynamic stream weighting for audio-visual speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1145–1157, 2012.

[17] M. Liu, Z. Xiong, S. M. Chu, Z. Zhang, and T. S. Huang, "Audio visual word spotting," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 785–788, IEEE, 2004.

[18] S. T. Shivappa, M. M. Trivedi, and B. D. Rao, "Hierarchical audio-visual cue integration framework for activity analysis in intelligent meeting rooms," in *Conference on Computer Vision and Pattern Recognition Workshops*, pp. 107–114, IEEE, 2009.

[19] H. Liu, T. Fan, and P. Wu, "Audio-visual keyword spotting based on adaptive decision fusion under noisy conditions for human-robot interaction," in *International Conference on Robotics and Automation*, pp. 6644–6651, IEEE, 2014.

[20] Z. Zhou, X. Hong, G. Zhao, and M. Pietikainen, "A compact representation of visual speech data using latent variables," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 181–187, 2014.

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

[22] Z. Zhou, G. Zhao, and M. Pietikainen, "Towards a practical lipreading system," in *Conference on Computer Vision and Pattern Recognition*, pp. 137–144, IEEE, 2011.

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

[24] Y. Pei, T.-K. Kim, and H. Zha, "Unsupervised random forest manifold alignment for lipreading," in *International Conference on Computer Vision*, pp. 129–136, IEEE, 2013.

[25] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *European Conference on Computer Vision*, pp. 679–692, Springer, 2012.

[26] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 177–190, 2014.

[27] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *International Conference on Computer Vision*, pp. 1513–1520, IEEE, 2013.

[28] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in *Conference on Computer Vision and Pattern Recognition*, pp. 1078–1085, IEEE, 2010.

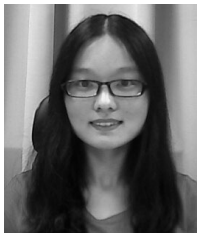
[29] O. Jesorsky, K. J. Kirchberg, and R. W. Frischholz, "Robust face detection using the hausdorff distance," in *Audio-and video-based biometric person authentication*, pp. 90–95, Springer, 2001.
[30] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Conference on Computer Vision and Pattern Recognition*, pp. 545–552, IEEE, 2011.
[31] L. Liang, R. Xiao, F. Wen, and J. Sun, "Face alignment via component-based discriminative search," in *European Conference on Computer Vision*, pp. 72–85, Springer, 2008.
[32] N. Duffy and D. Helmbold, "Boosting methods for regression," *Machine Learning*, vol. 47, no. 2-3, pp. 153–200, 2002.
[33] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Conference on Computer Vision and Pattern Recognition*, pp. 3360–3367, IEEE, 2010.
[34] S.-C. Zhu, C.-E. Guo, Y. Wang, and Z. Xu, "What are textons?," *International Journal of Computer Vision*, vol. 62, no. 1-2, pp. 121–143, 2005.
[35] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 2169–2178, IEEE, 2006.
[36] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Conference on Computer Vision and Pattern Recognition*, pp. 1794–1801, IEEE, 2009.
[37] S. Wan and J. Aggarwal, "Spontaneous facial expression recognition: A robust metric learning approach," *Pattern Recognition*, vol. 47, no. 5, pp. 1859–1868, 2014.
[38] A. Rogozan and P. Deléglise, "Adaptive fusion of acoustic and visual sources for automatic speech recognition," *Speech Communication*, vol. 26, no. 1, pp. 149–161, 1998.
[39] M. Tariquzzaman, S. M. Gyu, K. J. Young, N. S. You, and M. Rashid, "Performance improvement of audio-visual speech recognition with optimal reliability fusion," in *International Conference on Internet Computing & Information Services*, pp. 203–206, IEEE, 2011.
[40] S. Tamura, K. Iwano, and S. Furui, "A stream-weight optimization method for multi-stream hmms based on likelihood value normalization," in *International Conference on Acoustics, Speech, and Signal Processing*, pp. 469–472, IEEE, 2005.
[41] A. Adjoudani and C. Benoit, "On the integration of auditory and visual parameters in an hmm-based asr," in *Speechreading by humans and machines*, pp. 461–471, Springer, 1996.
[42] G. Potamianos and C. Neti, "Stream confidence estimation for audio-visual speech recognition," in *INTERSPEECH*, pp. 746–749, Citeseer, 2000.
[43] I. Matthews, J. A. Bangham, and S. Cox, "Audiovisual speech recognition using multiscale nonlinear image decomposition," in *International Conference on Spoken Language Processing*, vol. 1, pp. 38–41, IEEE, 1996.
[44] T. W. Lewis and D. M. Powers, "Sensor fusion weighting measures in audio-visual speech recognition," in *Australasian conference on Computer science*, vol. 26, pp. 305–314, 2004.
[45] R. C. Rose and D. B. Paul, "A hidden markov model based keyword recognition system," in *International Conference on Acoustics, Speech, and Signal Processing*, pp. 129–132, IEEE, 1990.
[46] M. Gurban and J.-P. Thiran, "Information theoretic feature extraction for audio-visual speech recognition," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4765–4776, 2009.
[47] N. Harte and E. Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.
[48] S. Young, P. Woodland, and W. Byrne, "Htk: Hidden markov model toolkit v3.4.1," 2009. Available at http://htk.eng.cam.ac.uk/.
[49] L. 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.
[50] J. Foote, S. J. Young, G. J. Jones, and K. S. Jones, "Unconstrained keyword spotting using phone lattices with application to spoken document retrieval," *Computer Speech & Language*, vol. 11, no. 3, pp. 207–224, 1997.
[51] X. Li and H. Liu, "Sound source localization for hri using foc-based time difference feature and spatial grid matching," *IEEE Transactions on Cybernetics*, vol. 43, no. 4, pp. 1199–1212, 2013.
[52] C. T. Ishi, S. Matsuda, T. Kanda, T. Jitsuhiro, H. Ishiguro, S. Nakamura, and N. Hagita, "A robust speech recognition system for communication robots in noisy environments," *IEEE Transactions on Robotics*, vol. 24, no. 3, pp. 759–763, 2008.

Pingping Wu received a B.E. degree in Information and Computing Science from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2009. Currently, she is working toward a Ph.D. degree at the School of Electronics Engineering and Computer Science (EE&CS), Peking University (PKU). Her current research interests are facial expression recognition, smile analysis, visual speech recognition. Related papers have been published on ICRA, ICIP, ICPR and ICASSP.

Prof. Hong Liu received a Ph.D. degree in mechanical electronics and automation in 1996, and serves as a full professor in the School of Electronics Engineering and Computer Science (EE&CS), Peking University (PKU). Prof. Liu has been selected as Chinese Innovation Leading Talent supported by "National High-level Talents Special Support Plan" since 2013.

He is also the Director of Open Lab on Human Robot Interaction, PKU, his research fields include computer vision and robotics, image processing, and pattern recognition. Dr. Liu has published more than 150 papers and gained Chinese National Aero-space Award, Wu Wenjun Award on Artificial Intelligence, Excellence Teaching Award, and Candidates of Top Ten Outstanding Professors in PKU. He is an IEEE member, vice president of Chinese Association for Artificial Intelligent (CAAI), and vice chair of Intelligent Robotics Society of CAAI. He has served as keynote speakers, co-chairs, session chairs, or PC members of many important international conferences, such as IEEE/RSJ IROS, IEEE ROBIO, IEEE SMC and IIHMSP, recently also serves as a reviewer for many international journals such as Pattern Recognition, IEEE Trans. on Signal Processing, and IEEE Trans. on PAMI.

Xiaofei Li received a Ph.D. degree in Electronics in 2013 from Peking University. He has worked since Feb. 2014 at the PERCEPTION team at INRIA Grenoble Rhône-Alpes as a post-doctoral fellow. His research interests include audio/speech signal processing, sound/speech recognition, sound source localization and audio-visual fusion. Related papers have been published on IEEE Transactions on Cybernetics, IROS, Interspeech and so on.



Ting Fan received a B.E. degree in Biomedical Engineering in 2011 from Xidian University, Xi'an, China. She is working toward an M.S. degree at Shenzhen Graduate School, Peking University (PKU). Her current research interests are speech recognition and audio-visual keyword spotting.

Xuewu Zhang received a B.E degree in Communication Engineering from North China Electric Power University in 2013. He is working toward an M.S. degree at Shenzhen Graduate School, Peking University (PKU). His research interests lie in facial expression recognition, smile analysis and video surveillance.