# Audio-visual Keyword Spotting for Mandarin Based on Discriminative Local Spatial-temporal Descriptors

Hong Liu, Ting Fan, Pingping Wu (Corresponding author)
Engineering Lab on Intelligent Perception for Internet of Things(ELIP)
Key Laboratory of Machine perception and Intelligence
Peking University, Shenzhen Graduate School, CHINA
Email: hongliu@pku.edu.cn, Email: fanting19900126@126.com, Email: wupingping@pku.edu.cn

***Abstract*—Although keyword spotting (KWS) technologies have been successfully applied to some applications, most KWS systems have a common problem of noise-robustness when applied to real-world environments. Audio-visual keyword spotting (AVKWS) using both acoustic and visual information is a solution to complementarily solve the problem. Most existing audio-visual speech recognition (AVSR) systems extract geometric features as visual features, which heavily rely on accurate and reliable detection and tracking of facial feature points. To avoid this defect of geometric features, an appearance-based discriminative local spatial-temporal descriptor (disCLBP-TOP) is proposed in this paper, which devotes to extracting robust and discriminative patterns of interest. Besides, a parallel two-step recognition based on both acoustic and visual keyword searching and re-scoring is conducted, which complementarily makes the best of two modalities under different noisy conditions. Adaptive weights for decision fusion are generated using a sigmoid function based on reliabilities of the two modalities, capable of adapting to various noisy conditions. Experiments show that our proposed parallel AVKWS strategy based on decision fusion significantly improves the noise robustness and attains better performance than feature fusion based audio-visual spotter. Additionally, disCLBP-TOP shows more competitive performance than CLBP-TOP.**

## I. INTRODUCTION

Keyword spotting (KWS) [1,2] deals with the identification of some specific words in unconstrained speech. Compared with continuous speech recognition (CSR) [3] that performs a complete transcription of an input utterance, KWS can deal with situations where various disfluencies and artifacts make the full-scale speech recognition difficult. Besides, without entire utterance to decode, KWS also leads to less time complexity. Therefore, KWS is more suitable for some specific applications such as dialogue systems and has been widely researched in the past decades.

Although KWS technologies have achieved significant progress and have been successfully applied to some well-defined applications, most KWS systems have the common problem of noise-robustness when applied to real-world environment with dramatically changing noises such as background noises, engine noises and other voice activities. Audio-visual keyword spotting (AVKWS) using both acoustic and visual information is a solution to complementarily solve the problem. Distinguished from acoustic speech, visual speech won't be affected by acoustic noise and thus can make up for the accuracy reduction of audio-only speech recognition in acoustically noisy conditions.

Audio-visual speech recognition (AVSR) has drawn wide attention due to the fact that audio-visual integration can enhance speech perception [4,5]. AVSR aims at improving ASR performance by sufficiently utilizing the visual information of vocal organs during the articulating process, especially in acoustically noisy conditions. Most research in this field focuses on audio-visual isolated word recognition and connected word recognition [6,7], but few works concern about AVKWS. Additionally, existing AVKWS systems [8,9] are primarily English oriented, pretty little attention is paid on AVKWS for mandarin. These motivate us to develop an audio-visual keyword spotter for Mandarin that can adapt to different noisy conditions.

For visual front-ends, an appearance-based feature named local spatiotemporal descriptors (LBP-TOP) was firstly proposed by Guoying Zhao et.al [10] for lipreading and achieved nice performance. In their later work [11], a new spatiotemporal local texture descriptor (CLBP-TOP) was proposed for differentiating spontaneous from posed facial expressions. However, CLBP-TOP contains some redundant information and leads to greater computational complexity. Therefore, an appearance-based visual feature named discriminative completed local binary pattern from three orthogonal planes (disCLBP-TOP) is proposed in our paper by picking out the most robust and discriminative patterns.

For audio-visual integration, there are two general fusion strategies: decision and feature fusion. Decision fusion is applied as modality integration strategy in this paper due to its advantages. Integrating weights are generated using a sigmoid function based on reliabilities of the two modalities, which can adapt to various noisy conditions. As to AVKWS strategy, a parallel two-step recognition is conducted to make the best of the two modalities under different noisy conditions. Fig.1 shows the general framework of our AVKWS system.

## II. VISUAL FEATURE EXTRACTION

For visual front-ends, the key point is to extract a discriminative feature vector of mouth movements. Generally, geometric features, appearance features and their combined features are extracted to represent visual information [12]. Geometry-based feature extraction commonly relies on accurate and reliable detection and tracking of fiducial points or extraction of lip contour, which may be significantly influenced by factors such as light conditions and head movement. Consequently, it's difficult to practice in real environments. Appearance-based feature extraction rises an alternative way to extract features directly from pixel-data instead of feature points,

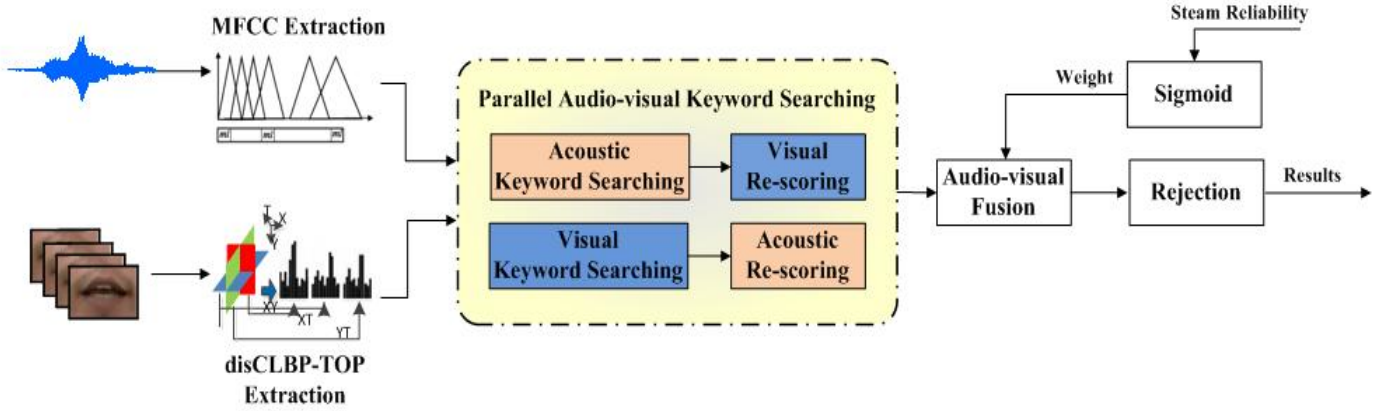


Fig. 1: Overall diagram of our audio-visual KWS system

which overcomes the drawbacks of geometry-based feature extraction.

However, most appearance-based visual speech recognition approaches consider features of lip or mouth regions in a global way, ignoring the local information that describes the local changes in space and time. Since the local information is of great significance, completed local binary pattern (CLBP) is proposed in [13] which extends local binary pattern (LBP) by adding the local difference of its central pixel intensity ($C$) and magnitude ($M$) besides sign ($S$). In [11], in order to derive dynamic information, the purely spatial CLBP was first extended to spatial-temporal domain by extracting CLBP features from three orthogonal planes (CLBP-TOP).

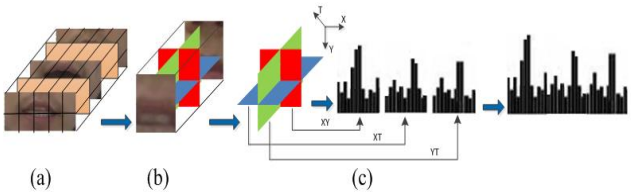


Fig. 2: Feature in each block volume (a) Block volumes (b) An exemplar block (c) CLBP features extracted from the exemplar block

Conventionally, CLBP-TOP of each block histograms are computed which are then concatenated to a single one to represent the appearance and motion of the mouth region sequence, as shown in Fig.2. However, applying CLBP-TOP directly to each divided block and then concatenating histograms into a single one may lead to another problem: the feature vector will be extremely long and the computational complexity will be substantially increased consequently. This motivates us to pick out the most representative and robust patterns to form the optimal subset of CLBP-TOP patterns. Inspired by the theory that the subset of effective patterns should be adaptively learnt from the database [14], we employ a learning model containing three layers to obtain the optimal subset of CLBP-TOP patterns [15]. Algorithm 1 shows the learning process of disCLBP-TOP.

Layer 1 searches more robust features with dominant pattern set, which is defined as the minimum set of pattern types covering $\delta$ ($0<\delta<1$) of all patterns. $p$ denotes the total number of pattern types in the $uth$ ( $u$=1: $XY$, 2: $XT$, 3: $YT$ plane ) and $P_{u,\xi}$ denotes the number of occurrences of pattern type $\xi$. The dominant pattern set of each orthogonal plane $J_u$ can be derived as following:

**Algorithm 1:** Learning process of disCLBP-TOP

**Input**: class $c$ with $n_c$ examples $\{S_1, S_2, ..., S_{n_c}\}$, B is the number of blocks
**Output**: extracted feature $J_{Global}$ of class $c$

**for** $n = 1$ **to** $n_c$ **do**
  divide $S_n$ into blocks $\{B_v | v = 1, ..., B\}$;
  compute sign dominant pattern $JS^n_{u,v}$ in $B_v$, $u = 1, 2, 3$;
  compute magnitude dominant pattern $JM^n_{u,v}$ in $B_v$, $u = 1, 2, 3$;
  $JS_{u,v} = JS_{u,v} \bigcap JS^n_{u,v}$;
  $JM_{u,v} = JM_{u,v} \bigcap JM^n_{u,v}$;
**end**
**for** $v = 1$ **to** $B$ *and* $u = 1$ **to** 3 **do**
  $JS_{Global} = JS_{Global} \bigcup JS_{u,v}$;
  $JM_{Global} = JM_{Global} \bigcup JM_{u,v}$;
**end**
$J_{Global} = JS_{Global} \bigcup JM_{Global}$;
**return** $J_{Global}$ ;

$$J_u = arg \min_{|J_u|} \left( \frac{\sum_{\xi \in J_u} P_{u,\xi}}{\sum_{k=1}^{p} P_{u,k}} \right) \geq \delta \quad (1)$$

where $|J_u|$ denotes the number of elements in $J_u$. In this way, the most frequently occurring patterns in each plane are preserved which tend to be reliable to represent the structure of the plane. The rarely occurring patterns are removed for they probably come from interference and may result in a sparse histogram.

Layer 2 ensures the discriminative power of features. In order to minimize the within-class scatter, it is desired that examples belonging to the same class have same patterns. Therefore, intersection of dominant pattern sets is carried out across all training examples in the same class. The optimal subset of CLBP-TOP patterns learned from class $c$ with $n_c$

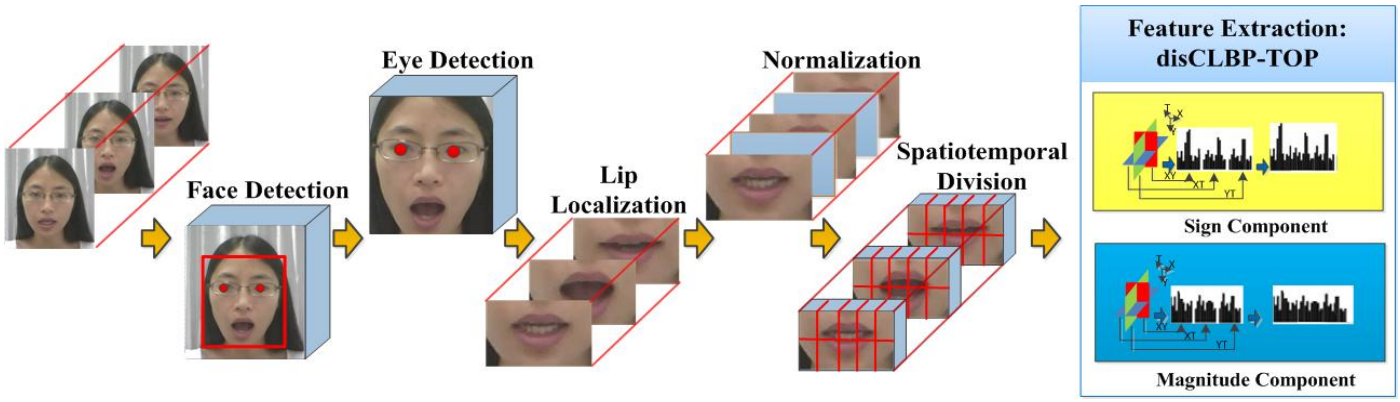


Fig. 3: The general framework of visual front end

examples can be expressed as:

$$J_c = \{\bigcup_{u=1}^{3}\bigcap_{n=1}^{n_c} JS_u^n\}\bigcup\{\bigcup_{u=1}^{3}\bigcap_{n=1}^{n_c} JM_u^n\} \quad (2)$$

where $JS_u^n$ and $JM_u^n$ denote the dominant pattern set from $uth$ plane of $nth$ example with respect to sign and magnitude component, separately. The central pixel intensity component is not considered here for it makes less contribution than the other two components [11,13].

Layer 3 constructs a global dominant patten set. Since disCLBP-TOP is a local descriptor, global location information is absent. To overcome the defect, the lip region sequence is usually equally divided into $B$ blocks in both spatial and temporal domain. And then the global feature is derived by concatenating $J_{c,v}$ together, which is extracted in block $B_v$, $v = 1,2,...,B$. The trained global feature from different classes is then put together as the reference for feature extraction of testing sets.

With the learning model, CLBP-TOP is optimized by seeking out dominant pattern sets and minimizing the within-class scatter. To the best knowledge of the authors, it is the first time CLBP-TOP being optimized and employed for audio-visual keyword spotting. The general framework of visual front end is illustrated in Fig.3.

## III. ADAPTIVE DECISION FUSION

For AVKWS, the way how acoustic and visual information is integrated significantly influences the final performance. Considering the integration level of acoustic and visual information, feature fusion and decision fusion are two broad integration strategies. Decision fusion is applied in our AVKWS system since it has some advantages over feature fusion in handling noisy conditions [16]: (1) Feature fusion needs more training data to ensure adequate probabilistic modeling. (2) Decision fusion can explicitly model the reliability of two modalities. (3) Integrating weights are relatively easy to generate using decision fusion since it independently handles the two modalities.

In this paper, conventional AVKWS based on HMM-garbage is adopted, where acoustic HMMs and visual HMMs are respectively trained. Adaptive integration is performed by linearly combining acoustic log-likelihoods and visual log-likelihoods of keyword candidates using the appropriate weights as follows [17]:

$$logP(O_{AV}|\lambda_i) = \gamma logP(O_A|\lambda_i^A) + (1-\gamma)logP(O_V|\lambda_i^V) \quad (3)$$

where $\gamma$ denotes the integration weight (0 to 1). $O_A$ and $O_V$ are the acoustic and visual feature sequences of a keyword candidate while $\lambda_i^A$ and $\lambda_i^V$ are the acoustic and visual HMM of keyword $i$. $logP(O_A|\lambda_i^A)$ and $logP(O_V|\lambda_i^V)$ represent the corresponding acoustic and visual log-likelihood.

In order to deal with various noise conditions, adaptive integration weights should be generated to combine the contributions of acoustic and visual modality. A number of reliability measures have been proposed in the literature. Referring to prior work [16], we select two reliability measures for each modality since they have the better recognition performance under diverse noisy conditions. The two reliability measures are the average difference against the maximum log-likelihood proposed by Neti [18] in Eq.(4) and variance of log-likelihoods proposed by Adjoudani [19] in Eq.(5).

$$D_1 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\max_j L_j - L_i\right) \quad (4)$$

$$D_2 = \frac{1}{N-1}\sum_{i=1}^{N}(L_i - \bar{L})^2 \quad (5)$$

where $L_i = logP(O|\lambda^i)$ is the output log-likelihood of the $i\bar{}th$ HMM, $\bar{L}$ is the mean log-likelihood.

Next, a sigmoid function is used to map the reliability measures to the stream weights since it is monotonic and bounded within zero and one. The mapping is defined as follows:

$$\gamma = \frac{1}{1+exp(-\sum_{i=1}^{4} w_i d_i)} \quad (6)$$

where $\mathbf{w} = [w_1, w_2, w_3, w_4]$ denotes the vector of sigmoid parameters and $\mathbf{d} = [D_{1,a}, D_{1,v}, D_{2,a}, D_{2,v}]$ denotes the reliability measure vector. Following prior work [20], the minimum classification error (MCE) approach is adopted to estimate the sigmoid parameters.

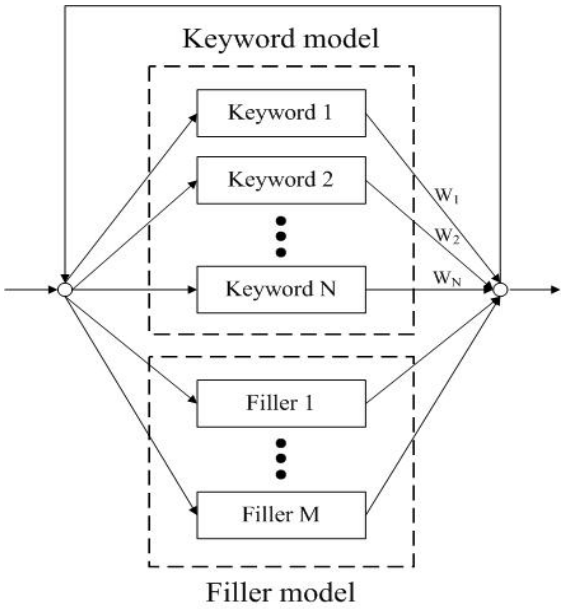


Fig. 4: Grammar of garbage based KWS

## IV. AUDIO-VISUAL KEYWORD SPOTTING STRATEGY

In this paper, conventional HMM-garbage based KWS [21] is applied, as depicted in Fig.4. Acoustic and visual keyword HMMs from left to right are trained based on whole word. Filler models are based on context independent phoneme or viseme, which is modeled with a 3-state HMM. Since the performance of conventional cascade strategy which performs visual re-scoring on the acoustic hypothesis in acoustic noisy conditions drops significantly, a parallel strategy is proposed to complementarily make the best use of two modalities as illustrated in Fig.5 [22].

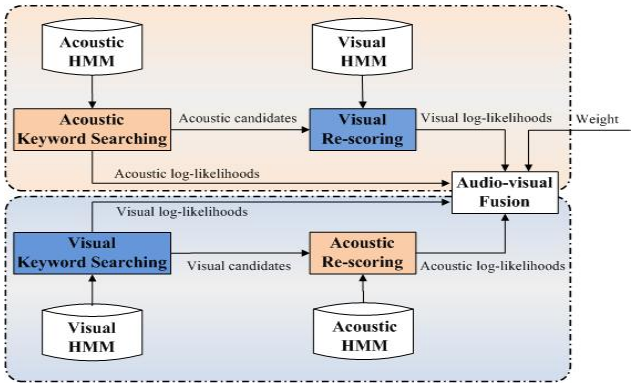


Fig. 5: Parallel audio-visual keyword spotting strategy

Acoustic keyword searching and visual keyword searching is first conducted in parallel, generating acoustic keyword candidates and visual keyword candidates with corresponding log-likelihoods. For a keyword candidate obtained by either modality, re-recognition based on the other modality of the keyword is then performed. For each acoustic and visual keyword candidate, an acoustic and a visual log-likelihood as well as corresponding reliability measure vector $[D_{1,a}, D_{1,v}, D_{2,a}, D_{2,v}]$ (Eq.(4) and Eq.(5)) can be obtained. With the reliability measure vector available, the integrating weight can be calculated using the sigmoid function. Then, integrated scores are generated by linearly combining the acoustic and visual log-likelihoods using the estimated weights (Eq.(3)).

In order to remove false alarms, rejection based on likelihood ratio [23] (or log likelihood difference) is conducted as follows:

$$logP(O_{AV}|\lambda_i, Filler) = logP(O_{AV}|\lambda_i) - logP(O_{AV}|Filler) \quad (7)$$

where $logP(O_{AV}|\lambda_i)$ and $logP(O_{AV}|Filler)$ denote the integrated log likelihood of keyword model $\lambda_i$ and filler model. 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 rejected.

For the remaining candidates after rejection, whether acoustic keyword candidate and visual keyword candidate are overlapped in time should be determined and specially handled, which may lead to higher detection rates. Therefore, a criterion carried out to deal with the overlapping situation. For each acoustic and visual keyword candidates, if the middle time point of one modality keyword candidate falls within the time region of the other modality keyword candidate, the candidate with greater integrated log-likelihoods is determined as a true keyword while the other is regarded as false alarm. For other cases, candidates are determined as true keywords.

## V. EXPERIMENTS AND DISCUSSIONS

### A. Experimental Setup

A new audio-visual database of mandrin recorded by 20 subjects (12 males and 8 females) is establish to conduct AVKWS experiments since existing audio-visual databases rarely concern about AVKWS of mandrin. Our database is collected in an acoustic quiet environment with controlled normal light conditions. The audio speech is recorded at the sampling rate of 16 kHz and 16 bits per sample. The video image is synchronously collected at 20 frames per second with a resolution of 640 $\times$ 480. each subject utters 300 sentences. We define 30 keywords frequently used in our task of human-robot interaction (HRI). The total duration is approximately 40 hours. Fig.6 shows some exemplar video frames in our database.

Commonly used Mel-frequency cepstral coefficients (MFCCs) and its delta as well as delta delta are extracted as the acoustic feature using HTK toolbox [24]. For visual preprocessing, faces and eyes are first detected automatically using the trained haar-cascade of OpenCV 2.4.6. According to the relative location of mouth and eyes, mouth region can be localized. Video normalization is performed using the approach in [25] in order to extract finer multi-resolution features in the following step. For visual feature, disCLBP-TOP is applied as stated in Section II. To illustrate, both acoustic and visual keyword HMMs are trained based on whole word, which are labeled by experts. The number of states of whole-word based keyword HMMs is in proportion to the number of the phonetic units in keywords. Figure of merit (FOM) is utilized as our performance measure.

Our database is divided into three sets to allow speaker-independent recognition: (1) 2100 clean utterances from 7

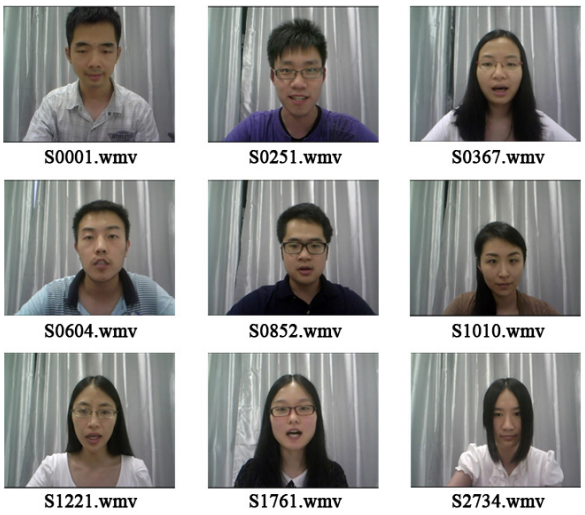


Fig. 6: Exemplar video frames in our database

subjects are used to train acoustic and visual HMMs. (2) $300 \times 6 \times 3 = 5400$ utterances from 6 subjects at various acoustic SNRs (artificially adding white noise at SNR of 20dB, 10dB and 0dB) are utilized to estimate the sigmoid parameters. (3) $300 \times 7 \times 2 \times 5 = 21000$ utterances from 7 subjects with different noises (white and babble noise) and different SNRs (20dB, 15dB, 10dB, 5dB and 0dB) are used to test the performance of our AVKWS system under various noise conditions.

### *B. Audio-visual Recognition*

Table I and II shows the performances of unimodal and bimodal KWS system respectively using CLBP-TOP and disCLBP-TOP under white and babble noise.

TABLE I: Audio-only, vision-only and audio-visual performances in terms of FOM using CLBP-TOP and disCLBP-TOP under white noise

| SNR(dB) | 20 | 15 | 10 | 5 | 0 |
|---|---|---|---|---|---|
| Audio | 74.7% | 57.3% | 39.4% | 18.6% | 6.4% |
| Vision-CLBP-TOP | 30.3% | 30.3% | 30.3% | 30.3% | 30.3% |
| Vision-disCLBP-TOP | 32.7% | 32.7% | 32.7% | 32.7% | 32.7% |
| AV-CLBP-TOP | 74.8% | 63.9% | 47.2% | 36.5% | 31.4% |
| AV-disCLBP-TOP | 75.1% | 65.2% | 50.3% | 38.1% | 34.6% |

TABLE II: Audio-only, vision-only and audio-visual performances in terms of FOM using CLBP-TOP and disCLBP-TOP under babble noise

| SNR(dB) | 20 | 15 | 10 | 5 | 0 |
|---|---|---|---|---|---|
| Audio | 70.0% | 49.8% | 37.4% | 17.2% | 6.1% |
| Vision-CLBP-TOP | 30.3% | 30.3% | 30.3% | 30.3% | 30.3% |
| Vision-disCLBP-TOP | 32.7% | 32.7% | 32.7% | 32.7% | 32.7% |
| AV-CLBP-TOP | 73.1% | 63.2% | 47.5% | 34.7% | 30.6% |
| AV-disCLBP-TOP | 74.8% | 64.2% | 51.9% | 37.4% | 34.3% |

We can observe that for both conditions, as the SNR decreases, the recognition performances of audio-only KWS system degrade significantly. While the performances of vision-only system remain the same since our database is collected under controlled normal light conditions. Besides, performances of bimodal KWS system outperforms the unimodal KWS system, owing to the complementary contribution of acoustic and visual modality. In addition, it can be observed that our approach works well for untrained noise conditions including different noise levels as well as noise types. Considering the vision-only and audio-visual performances using CLBP-TOP and disCLBP-TOP, we can see that the performance of our proposed disCLBP-TOP shows more competitive performance than CLBP-TOP proposed in [12] since disCLBP-TOP extracts the most robust and discriminative patterns of interest.

Then we compare our AVKWS system performance (decision fusion) and the audio-visual keyword spotter (feature fusion) proposed in [8] on our database. As shown in Table III, we can see that the performances of bimodal AVKWS systems using both decision fusion and feature fusion outperform unimodal systems from an overall point of view. While the integrated performance of the feature-based audio-visual keyword spotter in [8] is worse than vision-only at SNR of 0dB, our bimodal performance is at least equal to or better than that of unimodality. This phenomenon of the feature-level fusion approach can be explained that under extreme low SNR, the audio information introduces harmful cues and may degrade the overall performance of audio-visual fusion. For our decision-level fusion method, the contribution of each modality is combined using optimal weights adaptive to current noise conditions, which complementarily produces a better overall performance.

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% | 57.3% | 39.4% | 18.6% | 6.4% |
| Vision-only | 32.7% | 32.7% | 32.7% | 32.7% | 32.7% |
| Feature-level AV | 74.3% | 64.4% | 48.5% | 35.7% | 29.6% |
| Decision-level AV | 75.1% | 65.2% | 50.3% | 38.1% | 34.6% |

## VI. Conclusions

This paper develops an audio-visual keyword spotting (AVKWS) system for mandarin based on decision fusion that can adapt to various noise conditions. Instead of geometric features, an appearance-based visual feature disCLBP-TOP is proposed to extract the robust and discriminative patterns of interest. Appropriate weights are generated using a sigmoid function based on reliabilities of the two modalities to combine acoustic and visual contributions. As to AVKWS strategy, a parallel two-step recognition is conducted to make the best of the two modalities to obtain better performance under various noisy conditions. Experimental results show that bimodal performance outperforms unimodal performance especially in noisy conditions. Our AVKWS system based on adaptive decision fusion has a better performance the feature fusion based audio-visual keyword spotter. Additionally, since the proposed disCLBP-TOP can represent local information more accurately, performance using disCLBP-TOP outperforms that of approaches using CLBP-TOP.

## ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (NSFC, nos. 61340046, 60875050, 60675025), the National High Technology Research and Development Programme of China (863 Programme, no. 2006AA04Z247), the Scientific and Technical Innovation Commission of Shenzhen Municipality (nos. JCYJ20120614152234873, CXC201104210010A, JCYJ20130331144631730, JCYJ20130331144716089), and the Specialized Research Fund for the Doctoral Programme of Higher Education (SRFDP, no. 20130001110011).

## REFERENCES

[1] S. Zhang, Z. Shuang, Q. Shi, and Y. Qin, Improved mandarin keyword spotting using confusion garbage model. International Conference on Pattern Recognition (ICPR), pp. 3700-3703, 2010.

[2] H. Li, J. Han, and T. Zheng, Mandarin keyword spotting using syllable based confidence features and SVM, International Conference on Intelligent Control and Information Processing, vol. 1, pp. 256-259, 2011.

[3] A. A. Abdelhamid, W. H. Abdulla and B. A. MacDonald, WFST-based large vocabulary continuous speech decoder for service robots, Proceedings of the International Conference on Imaging and Signal Processing for Healthcare and Technology, pp.150-154, 2012.

[4] T. Yoshida, K.Nakadai, and H. G. Okuno, Automatic speech recognition improved by two-layered audio-visual integration for robot audition, IEEE/RAS International Conference on Humanoid Robots, pp. 604-609, 2009.

[5] 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.

[6] J. N. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes, DBN based multi-stream models for audio-visual speech recognition, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 993-996, 2004.

[7] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell, Visual speech recognition with loosely synchronized feature streams, International Conference on Computer Vision (ICCV), pp. 1424-1431, 2005.

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

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

[10] G. Zhao, and M. Pietikainen, Dynamic texture recognition using local binary patterns with an application to facial expressions, IEEE Transaction on Pattern Analysis Machine Intelligence, vol. 29, no. 6, pp. 915-928, 2007.

[11] T. Pfister, X. Li, G. Zhao, and M. Pietikainen, Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework, IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 868-875, 2011.

[12] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, Audio-visual automatic speech recognition: An overview, Issues in Visual and Audio-Visual Speech Processing, pp. 22-23, 2004.

[13] Zhenhua Guo, Lei Zhang, and David Zhang, A completed modeling of local binary pattern operator for texture classification, IEEE Transactions on Image Processing, vol. 19, no. 6, pp. 1657C1663, 2010.

[14] Y. Guo, G. Zhao, and M. Pietikainen, Discriminative features for texture description, Pattern Recognition, vol. 45, no. 10, pp. 3834-3843, 2012.

[15] P. P. Wu, H. Liu, and X. W. Zhang, Spontaneous versus posed smile recognition using discriminative local spatial-temporal descriptors, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2014, in press.

[16] J. S. Lee, and C. H. Park, Adaptive decision fusion for audio-visual speech recognition[J], Speech recognition, technologies and applications, pp. 550, 2008.

[17] A. Rogozan, and P. Delglise, Adaptive fusion of acoustic and visual sources for automatic speech recognition, Speech Communication, vol. 26, no. 1-2, pp. 149-161, 1998.

[18] G. Potamianos, and C. Neti, Stream confidence estimation for audio-visual speech recognition, Proceedings of the International Conference on Spoken Language Processing, pp. 746-749, 2000.

[19] A. Adjoudani, and C. Benoit, On the integration of auditory and visual parameters in an HMM-based ASR, Speechreading by Humans and Machines: Models, Systems, and Applications, pp. 461-472, 1996.

[20] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, Recent advances in the automatic recognition of audio-visual speech, Proceedings of the IEEE, vol. 91, no. 9, pp. 1306-1326, 2003.

[21] I. Szoke, P. Schwarz, P. Matejka, L.Burget, et al, Comparison of keyword spotting approaches for informal continuous speech, Proceedings of Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, 2005.

[22] H. Liu, T. Fan, and P. P. Wu, Audio-visual Keyword Spotting Based on Adaptive Decision Fusion under Noisy Conditions for Human-Robot Interaction, IEEE International Conference on Robotics and Automation (ICRA), 2014, in press.

[23] R. C. Rose, and D. B. Paul, A hidden Markov model based keyword recognition system, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 129-132, 1990.

[24] HTK, The Hidden Markov Model Toolkit (HTK), version 3.4.1 http://htk.eng.cam.ac.uk/, 2009.

[25] Ziheng Zhou, Guoying Zhao and M. Pietikainen, Towards a Practical Lipreading System, International Conference on Computer Vison and Pattern Recogntion, pp. 137-144, 2011