2014 IEEE International Conference on Robotics & Automation (ICRA)
Hong Kong Convention and Exhibition Center
May 31 - June 7, 2014. Hong Kong, China

Audio-visual Keyword Spotting Based on Adaptive Decision Fusion under Noisy Conditions for Human-Robot Interaction

Hong Liu, Ting Fan and Pingping Wu*

Abstract— **Keyword spotting (KWS) deals with the identification of keywords in unconstrained speech, which is a natural, straightforward and friendly way for human-robot interaction (HRI). Most keyword spotters have the common problem of noise-robustness when applied to real-world environment with dramatically changing noises. Since visual information won't be affected by the acoustic noise, it can be utilized to complementarily improve the noise-robustness. In this paper, a novel audio-visual keyword spotting approach based on adaptive decision fusion under noisy conditions is proposed. In order to accurately represent the appearance and movement of mouth region, an improved local binary pattern from three orthogonal planes (ILBP-TOP) is proposed. Besides, a parallel two-step recognition based on acoustic and visual keyword candidates is conducted and generates corresponding acoustic and visual scores for each keyword candidate. Optimal weights for combining acoustic and visual contributions under diverse noise conditions are generated using a neural network based on reliabilities of the two modalities. Experiments show that our proposed audio-visual keyword spotting based on decision fusion significantly improves the noise robustness and attains better performance than feature fusion based audio-visual spotter. Additionally, ILBP-TOP shows more competitive performance than LBP-TOP.**

I. INTRODUCTION

Automatic speech recognition (ASR) [1,2] is a natural, effective and friendly way for human-robot interaction (HRI) and has been widely researched in the past decades. In some scenarios of HRI, continuous speech recognition (CSR) [3] that performs a complete transcription of the input utterance is not necessary since the key information lies in only part of the input utterance. In such scenarios, keyword spotting (KWS) [4,5] can obtain fast access to information since it merely identifies the occurrences of some predefined words instead of the whole utterance. Compared with CSR, 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 concerning our task of HRI. For KWS, there are three typical approaches: HMM-garbage based acoustic KWS, phoneme lattice based KWS and large vocabulary continuous speech recognition (LVCSR)-based KWS [6]. For our task of HRI focusing on a certain domain, HMM-garbage based KWS is a good choice due to its timeliness.

Although current KWS systems for HRI have a satisfying recognition accuracy in relatively quiet environments, their performances degrade significantly when applied to real-world environment with dramatically changing noises such as background noises, robots' own noises and other voice activities. Distinguished from acoustic signal, visual signal won't be affected by the acoustic noise and can benefit human speech perception. As to a robot with eyes (cameras) and ears (microphones) available for face-to-face HRI, visual information can be extracted synchronously and combined together with auditory information to improve the noise robustness of a keyword spotter.

Motivated by the contribution of visual information in ASR, many researches have been done on audio-visual speech recognition (AVSR) [7,8]. Generally, there are two broad audio-visual information fusion categories: feature fusion and decision fusion. While many researches have been conducted on AVSR, few works concern about audio-visual keyword spotting (AVKWS). Ming Liu etc. designed an audio-visual word spotter which adopts the feature fusion by directly concatenating acoustic feature and visual feature to a single larger feature vector without any appropriate transformations according to various noisy conditions [9]. Shankar T. Shivappa proposed a hierarchical audio-visual cue integration framework for activity analysis in intelligent meeting rooms where KWS is merely a small task used to evaluate the performance of beamforming, without specific research on AVKWS [10]. Additionally, pretty little attention is paid on AVKWS for Mandarin.

These motivate us to propose an audio-visual keyword spotting approach for Mandarin that can adapt to different noisy conditions. In our paper, an appearance-based visual feature ILBP-TOP is proposed which considers the changes of a mouth region both in space and time. By comparing different fusion strategies, a decision fusion instead of feature fusion is applied to complementarily combine the two modalities. Appropriate weights for combining acoustic and visual contributions are generated using a neural network based on stream reliabilities. For AVKWS, a parallel two-step recognition is conducted to complementarily combine the two modalities so as to obtain better performance under various noisy conditions. Finally, an additional step is taken

This work is supported by National Natural Science Foundation of China (NSFC, No.61340046, 60875050, 60675025), National High Technology Research and Development Program of China (863 Program, No.2006AA04Z247), Scientific and Technical Innovation Commission of Shenzhen Municipality (No. JCYJ20120614152234873, CXC201104210010A, JCYJ20130331144631730, JCYJ20130331144716089).

H. Liu is with the Engineering Lab on Intelligent Perception for Internet of Things(ELIP) and the Key Laboratory of Machine perception and Intelligence, Peking University, Beijing, 100871 CHINA. *hongliu@pku.edu.cn*

T. Fan is with the Engineering Lab on Intelligent Perception for Internet of Things(ELIP), Shenzhen Graduate School of Peking University, Shenzhen, 518055 CHINA. *fanting*19900126@126.*com*

Corresponding author P. Wu is with the Engineering Lab on Intelligent Perception for Internet of Things(ELIP), Shenzhen Graduate School of Peking University, Shenzhen, 518055 CHINA. *wupingping@pku.edu.cn*

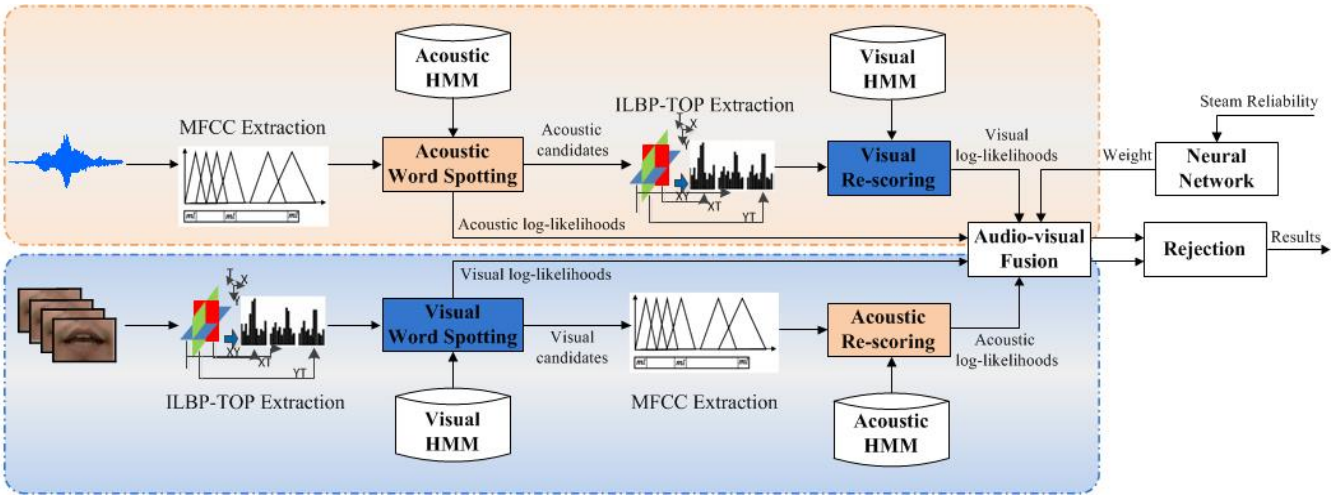

Fig. 1: The overall diagram of our audio-visual KWS system

to deal with the acoustic and visual keyword candidates that overlap in time. The overall diagram of our KWS system is shown in Fig.1.

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 [11]. Geometry-based feature extraction commonly relies on accurate and reliable detection and tracking of fiducial points or extraction of lip contour [9,12], which may be significantly influenced by factors such as light conditions and head movement. Consequently, it's difficult to practice in the real environments. Appearance-based feature extraction rises an alternative way to extract features directly from pixel-data instead of feature points, 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 [13,14]. Since the local information is of great significance, an improved local binary pattern from three orthogonal planes (ILBP) is proposed, which considers both changes in time and space of a mouth region. Moreover, a multi-resolution method is applied and Adaboost is also presented to perform feature selection.

A. Improved local binary pattern

In order to obtain the local information, a structure kernel with size of $K \times K$ is considered. There exists 2^{K^2} patterns while only $2^{K^2}-1$ are valid since the kernel with all elements 0 and that with all 1 convey the same information. Some examples of a 3×3 structure kernel are shown in Fig.2. As it is shown, binary patterns of the structure kernel can represent primitive texture like lines, spots and corners.

Local binary pattern (LBP) is a typical local feature descriptor, which has shown excellent performance in texture classification [15]. For each pixel in an image, a unique P-bit

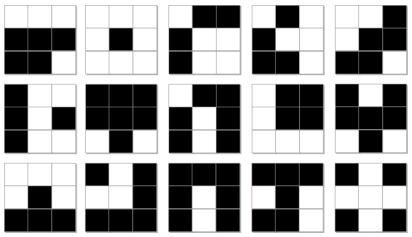

Fig. 2: A randomly chosen subset of 15 out of 511 possible local structure kernels within a 3×3 neighborhood

pattern code is generated by assigning binomial coefficient 2^P to each sign $s(g_p - g_c)$ as follows:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c)2^p, s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (1)$$

where P denotes the number of pixels in the surrounding neighborhood and R denotes the circle radius. g_c represents the gray value of the central pixel (x_c, y_c) of the local neighborhood and g_p denotes the gray values of P equally spaced pixels on the circle with radius R.

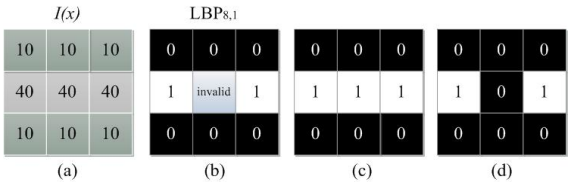

Fig. 3: Results of a micro image with and without coding the central pixel: (a) $I(x)$ is a micro image consisting of 3×3 pixels. (b)shows the result of $LBP_{8,1}$ without coding the central pixel, (c) and (d) are the two possible patterns performing coding the central pixel

For LBP, only a subset of $2^8 - 1 = 255$ of all $2^9 - 1 = 511$ patterns is available concerning a 3×3 structure. For a micro image shown in Fig.3(a), the representation result of $LBP_{8,1}$ is shown in Fig.3(b). As the central pixel is not coded, the describing of the micro image is inaccurate. This motivates us to make an modification of the original LBP by assigning the central pixel a specific binary code. Let $\bar{g}$ be the intensity mean of the kernel structure, that is,

$$\bar{g} = \frac{1}{p+1}\left(\sum_{p=0}^{p-1} g_p + g_c\right) \quad (2)$$

Then we reformulate Eq.(1) and rewrite the ILBP as

$$ILBP_{P,R} = \sum_{p=0}^{P} s(g_p - \bar{g})2^p, s(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (3)$$

When representing the micro image in Fig.(a), ILBP derives the right pattern. The coding sequence is shown in Fig.4(a) when $P = 8$ and $R = 1$. As P, R can be positive integer, Different resolutions are available for ILBP as shown in Fig.4(b)

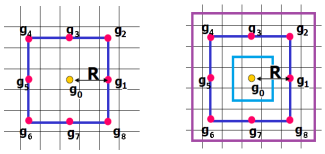

Fig. 4: (a) corresponding number arrangement (b) different resolution ILBP

B. ILBP from three orthogonal planes

Though ILBP can obtain more accurate pattern representing, it only extract the appearance feature in spatial domain. However, for lipreading, it is a key step to extract not only the static information from each frame but also the dynamic information among frames. In the work of [30], LBP was extended from spatial domain to spatiotemporal domain by extracting LBP features from three orthogonal planes (XY, XT, YT) which is named LBP-TOP. In order to derive the dynamic information of utterance process, our ILBP is also extended to the spatiotemporal domain by extracting ILBP features from three orthogonal planes (ILBP-TOP).

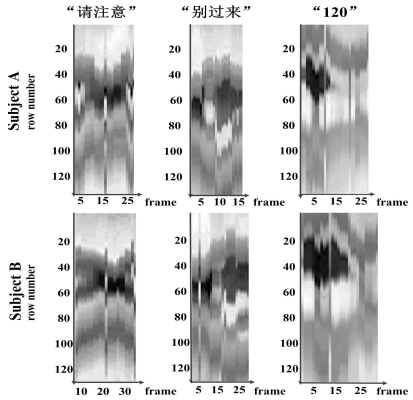

Fig. 5: Temporal patterns of the video of three keywords spoken by two subjects in Chinese. The abscissa and the ordinate show the image frame and row indices respectively.

LBP-TOP has the ability to describe spatiotemporal signals and is robust to monotonic gray-scale changes which has been successfully applied in dynamic feature recognition [16]. ILBP-TOP keeps all the advantages and additionally gives more accurate description of the local image structure. As the utterance process of different keywords could be regarded as different dynamic textures, ILBP-TOP is utilized to represent the mouth movement. As can be seen in Fig.5, although with different temporal resolutions, the patterns of the same keyword show similarities and are distinguishable from others in the meantime.

In order to obtain the location information, the video sequence of the mouth region is considered as a volume which is divided into $B \times C \times D$ blocks in spatial-temporal domain. The orthogonal planes in the block are called slices. In each block, ILBP-TOP 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.6.

A ILBP-TOP histogram of the mouth movements can be defined as

$$H_{m,n,l,j,i} = \sum_{x,y,t} T\{ILBP^j_{P,R}(x,y,t) = i\}, \quad (4)$$

$$i = 0, \cdots, \omega_j - 1; j = 0, 1, 2$$

where ω_j is the number of bins produced by $ILBP$ operator in the jth plane ($j = 0 : XY, 1 : XT, 2 : YT$) and $ILBP^j_{P,R}(x,y,t)$ represents the $ILBP$ code of central pixel (x,y,t) in the jth plane. m is the row index, n is the column index and l is the utterance length.

$$T\{A\} = \begin{cases} 1, if\ A\ is\ true \\ 0, if\ A\ is\ false \end{cases} \quad (5)$$

The histograms must be normalized to get a coherent representation:

$$H_{m,n,l,j,i} = \frac{H_{m,n,l,j,i}}{\sum_{k=0}^{2^P-1} H_{m,n,l,j,k}} \quad (6)$$

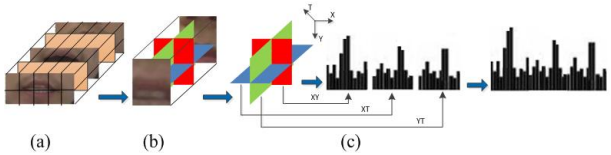

Fig. 6: Feature in each block volume. (a) Block volumes (b) ILBP features from three orthogonal planes. (c) Concatenated features for one block volume with the appearance and motion

C. Feature selection based on Adaboost

From the work of [16], using a combination of different resolution features (P, R taking different values, such as $LBP_{8,3}$, $LBP_{4,1}$) with different block volume sizes ($B \times C \times D$, B, C, D taking different values) could improve the performance. However, if all the different features were directly concatenated, the feature vector would be too long. Using them straightly will lead to high computational complexity and memory requirements. As it is known, Adaboost shows good performance not only in feature classification but also in feature selection. There are two strategies while using Adaboost for feature selection: block-based method and slice-based method. The block-based selection only considers the location information, ignoring the principal appearance and motion. Besides, if one block is selected, three slices in the block are all included. The slice-based selection devotes to find out which slice (XY, XT, YT) gives more contribution while also considering the location information, which further analysis features. Thus slice-based selection is adopted in our work to select effective features from a large feature set of different resolutions from different block volume sizes.

Generally in selecting, Adaboost treat each slice histogram feature as a weak classifier and picks up the best one of these classifiers and boosts the weights on the error examples. The

next filter is selected which gives the best performance on the errors of the previous one. After T rounds of iteration, T features are selected out. The detailed steps of attribute sorting with Adaboost algorithm could be seen in [17]. The general framework of visual front end is illustrated in Fig.7.

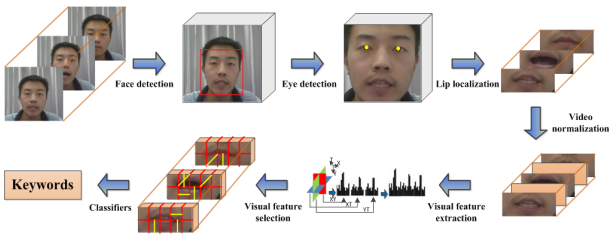

Fig. 7: The general framework of visual front end

III. ADAPTIVE AUDIO-VISUAL INTEGRATION

For AVKWS, integration module is the main issue that determines the performance. Obviously, contributions of acoustic and visual information are different under various noisy conditions. Therefore, the way where and how audio and visual information is integrated significantly influences the final performance. Generally, there are two broad fusion categories: feature fusion and decision fusion [18]. Feature fusion directly concatenates the features of the two modalities into a larger feature vector in a plain way or adopting some appropriate transformations. Recognition is then conducted using a single classifier based on the composite feature vector. Contrarily, for decision fusion, two individual classifiers are adopted, one is audio based and the other is vision based. Final results are obtained by combining the outputs of the two modalities.

Compared with feature fusion, decision fusion approaches have some advantages in handling noisy conditions [11,18]: (1) Feature fusion concatenates acoustic and visual features into a larger feature vector with higher dimensionality, thus more training data are needed to ensure adequate probabilistic modeling. (2) In contrast to feature fusion, decision fusion can explicitly model the reliability of two modalities, which is of great significance since discrimination power of the two modalities may vary widely. (3) According to different noisy conditions, integrating weights are relatively easy to generate in order to adaptively control the contributions of the two modalities using decision fusion which independently handles the two modalities.

As to our task of developing a noise-robust AVKWS system, decision fusion approach is the better choice due to advantages stated above. Similar to CSR, there are three temporal levels to integrate modality likelihoods: "Early" integration, "Late" integration and "Intermediate" integration [18]. "Late" integration is applied in our paper since state-level or word-level asynchrony is allowed. Besides, conventional AVKWS based on HMM is utilized, where acoustic HMMs and visual HMMs are respectively trained and provide corresponding modality likelihoods of a given multimedia source (acoustic and visual speech). Integrated scores can be obtained by linearly combining acoustic and visual log-likelihoods of keyword candidates using the appropriate weights as follows [19]:

$$logP(O_{AV}|\lambda_i) = \gamma logP(O_A|\lambda_i^A) + (1-\gamma) logP(O_V|\lambda_i^V) \quad (7)$$

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. $logP(O_A|\lambda_i^A)$ and $logP(O_V|\lambda_i^V)$ represent the corresponding acoustic and visual log-likelihood.

In a specific environment, integration weight with a constant value can be estimated according to the optimal performance under such conditions. However, when the audio-visual environment changes dramatically, fixed integration weight is not sufficient to cope with the noise-varying conditions. Confronted with diverse noisy conditions, the main issue lies in obtaining adaptive integration weights based on the reliabilities of two modalities [20,21].

The reliability measure based on the output log-likelihoods of acoustic and visual HMMs has been commonly used. Taking the varying acoustic conditions for example, output log-likelihoods of each HMM in quiet environment show great differences while small differences are obtained when the environment is noisy. Large difference reflects less ambiguity and larger certainty for recognition [18].

Various forms of reliability measure have been implemented by researchers. Adjoudani and Benoit introduced a reliability measure based on the variance of log-likelihoods in each modality [22]:

$$D = \frac{1}{N-1}\sum_{i=1}^{N}(L_i - \bar{L})^2 \quad (8)$$

where $L_i = logP(O|\lambda^i)$ is the output log-likelihood of the i-th HMM, $\bar{L}$ is the mean log-likelihood of N HMMs. Potamianos and Neti proposed the differences between each pair of N-best hypotheses [23]:

$$D = \frac{2}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(L_i - L_j) \quad (9)$$

The average difference against the best hypothesis with the maximum log-likelihood is also calculated as reliability measure in Potamianos and Neti's work [23]:

$$D = \frac{1}{N-1}\sum_{i=1}^{N}\left(\max_j L_j - L_i\right) \quad (10)$$

Matthews et al. used inverse entropy of posterior probabilities since the entropy can represent the differences between the posterior probabilities [24]:

$$D = [-\frac{1}{N}\sum_{i=1}^{N}P(C_i|O)logP(C_i|O)]^{-1} \quad (11)$$

where $P(C_i|O)$ is the posterior probability.

Studies show that the average difference against the maximum log-likelihood (Diffmax) in Eq.10 has the best recognition performance under different noisy conditions from an overall point of view [18]. Lewis and Powers pointed

out that the intrinsic errors of other dispersion forms in measuring reliabilities lead to their inferiority to Diffmax [25]. Therefore, the reliability measure of average difference against the maximum log-likelihood is used in our paper.

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

$$\gamma = f(D_A, D_V) \tag{12}$$

In order to obtain adaptive weights for various noisy conditions, both clean and noise-corrupted speeches are utilized to train the neural network. Practically, noise-corrupted speeches are generated by artificially corrupting clean speeches with additive white noise at several SNRs (20dB, 10dB and 0dB) for training. The trained neural network can generate optimal weight of a keyword candidate for various noisy conditions, not limited to the noisy conditions used for training.

Detailed 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.

IV. KEYWORD SPOTTING STRATEGY

Aiming at investigating the benefit of visual modality to KWS using the adaptive weights, conventional HMM-garbage based KWS is applied, which is primarily used in the application fields such as dialogue system, command control and information consultation.

In our AVKWS system, acoustic and visual keyword HMMs from left to right are trained based on whole word since keyword dependence can improve the performance. Context independent phonemes and visemes are chosen as sub-word units and used as corresponding filler models. Each sub-word unit is modeled with a 3-state HMM and each state contains eight Gaussian components. Multiple occurences of keywords are allowed in our system. 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 the acoustic recognition performance in acoustic noisy conditions drops significantly, the simple strategy which merely performs visual re-scoring on the acoustic candidate is abandoned. A parallel strategy is proposed to complementarily make the best use of two modalities. Time synchrony of acoustic and visual signal should be ensured. With the trained acoustic and visual keyword HMMs as well as filler models available, acoustic and visual keyword searching is first conducted in parallel on the testing speech, generating a number of acoustic keyword candidates as well as visual keyword candidates with corresponding log-likelihoods. Not surprisingly, the acoustic keyword candidates are not necessarily the same as the visual ones, especially when the acoustic environment is too noisy. For a keyword candidate obtained by either modality with starting and ending time, re-scoring based on the other modality of the keyword is then performed. Therefore, each candidate obtains an acoustic and a visual log-likelihood. Additionally, the corresponding acoustic reliability D_A and visual reliability D_V can be calculated based on the candidate. 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 using the estimated weights. The general process is illustrated in Fig.1.

Finally, rejection is conducted based on the integrated scores of each keyword candidate to remove false alarms. Taking the noisy background into consideration, rejection based on likelihood ratio [26] (or log likelihood difference) is utilized due to its robustness to noise, as follows:

$$\log P(O_{AV}|\lambda_i, Filler) = \log P(O_{AV}|\lambda_i) - \log P(O_{AV}|Filler) \tag{13}$$

where $\log P(O_{AV}|\lambda_i)$ is the integrated log likelihood of keyword model λ_i and $\log P(O_{AV}|Filler)$ is the integrated log likelihood of filler model. Similarly, $\log P(O_{AV}|Filler)$ can be calculated based on the filler models 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 rejected.

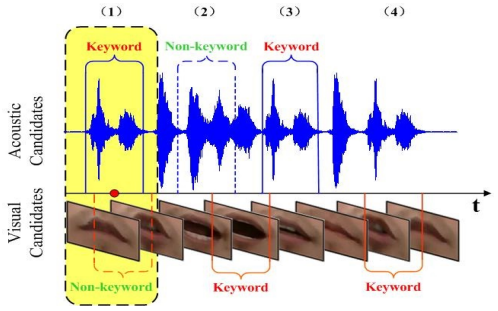

Fig. 8: Additional step to deal with the overlapping acoustic and visual candidates

Conventionally, recognition result analysis is performed by comparing it with the artificially labeled reference after keyword verification. As depicted in Fig.9, 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, we carry out a criterion to deal with the situation. For each acoustic and visual keyword candidates with corresponding time region and integrated loglikelihoods, if the middle time point of one modality keyword candidate falls within the time region of the other modality keyword candidate, it's regarded as the overlapping situation. Therefore, the candidate with greater integrated loglikelihood is determined as a true keyword while the other

is regarded as false alarm (case (1) in Fig.8). For other cases (acoustic and visual keyword candidates do not overlap in time), candidates are directly determined as true keywords (case (3),(4) in Fig.8).

V. Experiments and discussions

A. Experimental Setup

Distinguished from audio-only corpora, audio-visual databases need to take some additional challenges into consideration such as audio-visual data collection and storage. Therefore, only a few common databases are available for AVSR[16]. In addition, the existing audio-visual databases rarely concern about AVKWS of mandrin. Thus, we establish a new audio-visual database to conduct experiments considering AVKWS of mandrin.

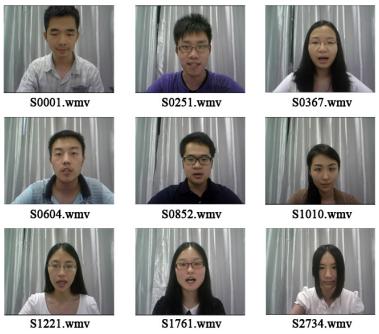

Fig. 9: Video images in our database

Our database is collected in an acoustic quiet environment under controlled normal light conditions. The video image is collected at 20 frames per second with a resolution of 640×480 by a video camera. The audio speech is synchronously recorded at the sampling rate of 16 kHz and 16 bits per sample. Fig.9 shows some exemplar video frames in our database. Our database contains 20 subjects (12 males and 8 females) and there are 300 utterances for each subject. We define 30 keywords that are frequently used in our task of HRI. The total duration of our audio-visual database is approximately 40 hours. Our AVKWS system will be applied on our Pengpeng II HRI oriented mobile robot system as depicted in Fig.10.

For acoustic feature, commonly used Mel-frequency cepstral coefficients (MFCCs) and its delta as well as delta delta are extracted using HTK toolbox [27]. On the other hand, ILBP-TOP is applied for visual feature extraction. 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 whole word [28], which are labeled by expert. The number of states of whole-word based keyword HMMs is in proportion to the number of the phonetic units in the keyword. Experimentally, the number of hidden neurons of the neural network is set to six and figure of merit (FOM) is utilized as our performance measure.

Our database is divided into three sets to allow speaker-independent recognition: (1) Training sets composed of 2100 clean utterances from 7 subjects are used to train all acoustic and visual HMMs. (2) Held-out sets consisting of $6 \times 300 \times 3 = 5400$ utterances from 6 subjects at various acoustic SNRs (artificially adding white noise at SNR of 20dB, 10dB and 0dB) are utilized to train the neural network for estimating the optimal weight used in decision fusion step. (3) Test sets ($7 \times 300 \times 2 \times 5 = 21000$ utterances from 7 subjects with different SNRs) are used to evaluate the performance of our AVKWS system under various noise conditions, which can be obtained by artificially corrupting the clean speech data with different noises (white and babble noise) at diverse SNRs (20dB, 15dB, 10dB, 5dB and 0dB).

Fig. 10: Pengpeng II HRI oriented mobile robot

B. Vision-only Recognition

For 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 [29] in order to extract finer multi-resolution features in the following step. Comparison experiments are conducted to explore the influence of dividing methods of block volume sizes. Besides, the impact of video normalization is also investigated. Since LBP-TOP features with $P = 8, R = 3$ and $P = 4, R = 1$ are widely used, ILBP-Top$_{8,3}$ and ILBP-Top$_{4,1}$ are utilized in our experiments. As shown in Table I, it can be observed that

TABLE I: Comparison of recognition results (FOM) of multi-resolution features with and without normalization

Blocks ($M \times N \times L$)	ILBP-Top$_{8,3}$ (%)		ILBP-Top$_{4,1}$ (%)	
	un-norm	norm	un-norm	norm
$5 \times 2 \times 3$ (90 slices)	23.6	28.2	20.5	25.7
$5 \times 2 \times 1$ (30 slices)	12.4	16.3	10.2	15.8
$5 \times 1 \times 3$ (45 slices)	21.1	23.2	18.3	21.6
$4 \times 2 \times 3$ (72 slices)	21.8	27.3	18.8	22.9
$4 \times 2 \times 1$ (24 slices)	10.7	14.4	8.5	14.8
$4 \times 1 \times 3$ (36 slices)	13.1	16.4	11.7	11.6

for both ILBP-Top$_{8,3}$ and ILBP-Top$_{4,1}$ with different block volume sizes, the feature extracted after video normalization generally has a more competitive performance than that without normalization. This consists with the work [29] and verifies the necessity of video normalization. Additionally, for both ILBP-Top$_{8,3}$ and ILBP-Top$_{4,1}$, a dividing method of $5 \times 2 \times 3$ results in the best performance. Moreover, ILBP-Top$_{8,3}$ with block volume divided into $5 \times 2 \times 3$ derives the

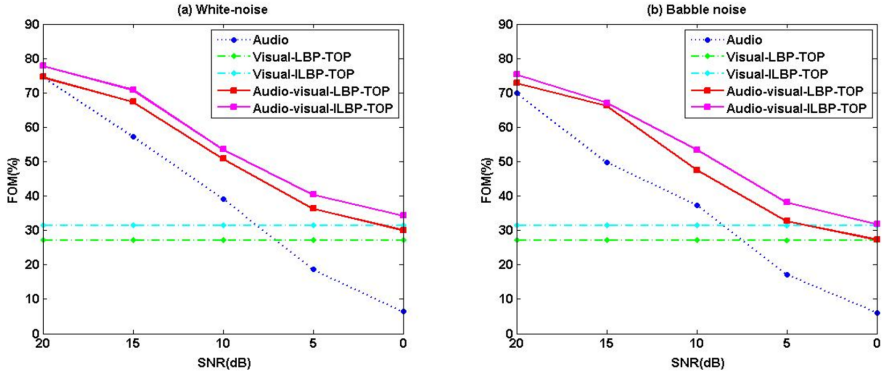

Fig. 11: Recognition performance of the audio-only, vision-only and audio-visual KWS system under white noisy and babble noisy conditions. (a) performance of white noisy condition using LBP-TOP and ILBP-TOP. (b)performance of babble noisy condition using LBP-TOP and ILBP-TOP.

most appealing results.

As presented in Section II, the multi-resolution features with the greatest contribution to performance are then selected using both slice-based and block-based Adaboost. Fig.11 illustrates the recognition performances of multi-resolution features with different number of slices selected respectively using slice-based and block-based Adaboost. We can see that the multi-resolution features selected using slice-based result in better performance than block-based Adaboost. Besides, for slice-based Adaboost, when 90 slices are selected from the total 594 slices ($(90+30+45+72+24+36)\times 2 = 594$), the best FOM reaches 31.4%, which is much better than that of ILBP-TOP$_{8,3}$ with 90 slices (28.2%) shown in Table I. Therefore, selecting features with the greatest contribution from different block sizes as well as different resolution does improve the recognition performance.

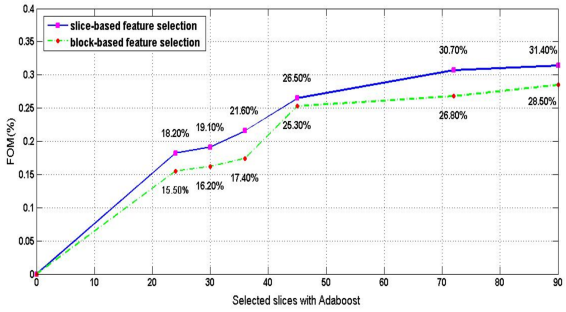

Fig. 12: Comparison performance of block-based and slice-based feature selection using Adaboost

C. Audio-visual Recognition

Fig.12 shows the recognition performances of the audio-only, vision-only and audio-visual KWS system using LBP-TOP and ILBP-TOP under various noise conditions(white noise and babble noise). It can be observed that the performance of audio-only recognition significantly degrades as the speech becomes noisier. The 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. Take the white-noise-ILBP-TOP case for example, FOM increases from 39.2% to 53.6% when SNR=10dB. Besides, clean speeches corrupted by white noise with SNR of 20dB, 10dB and 0dB are used to train the neural network and tests conducted on white noisy and babble noisy speeches at various SNR conditions (20dB, 15dB, 10dB, 5dB and 0dB) show that our approach also works well for untrained noise conditions including different noise levels as well as noise types. Considering the vision-only and audio-visual performances using LBP-TOP and ILBP-TOP for both white and babble noise conditions, we can observe that the performance of our proposed ILBP-TOP outperforms that of LBP-TOP proposed in [30].

TABLE II: 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	31.4%	31.4%	31.4%	31.4%	31.4%
Feature-level AV	76.3%	66.4%	49.2%	35.7%	26.6%
Decision-level AV	77.8%	70.9%	53.6%	40.4%	34.3%

Then we compare our AVKWS performance based on decision-level fusion using adaptive weights (using ILBP-TOP) with the feature-based audio-visual keyword spotter proposed in [9] on our database (white noise corrupted speech). As shown in Table II, it can be seen that our approach is more robust to noise than that of [9]. Our integrated audio-visual performance is at least equal to or better than that of unimodality while the integrated performance in [9] gets worse compared to vision-only performance at SNR of 0dB. 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. Contrarily, the contribution of acoustic and visual modality is combined using adaptive weights according to current SNR conditions in our decision-level fusion method, which may complementarily produce a better overall performance.

VI. Conclusions and Future Work

In order to obtain more effective HRI under various noisy conditions, this paper develops an audio-visual keyword spotter based on decision-level fusion for Mandarin that can adapt to different noise conditions. An appearance-based visual feature ILBP-TOP is proposed which considers the changes of a mouth region both in space and time and describes the local information more accurately. A neural network that takes acoustic and visual reliabilities as input

is trained to obtain the optimal weights under current noisy condition. The average difference against the maximum log-likelihood is adopted as the reliability measure in this work. As to AVKWS, a parallel two-step recognition based on both acoustic and visual modality is conducted in order to make the best use of the two modalities under various conditions. Integrated log-likelihoods can be obtained by linearly combining the log-likelihoods of two streams using the adaptive weights of certain noisy condition. Besides, an additional step is taken to deal with the overlapping acoustic and visual keyword candidates. Experimental results show that our audio-visual integration based on decision level does improve the noise robustness of the keyword spotter, which outperforms the feature fusion based audio-visual keyword spotter. The neural network works well for untrained noisy conditions including different noise levels as well as noise types. Additionally, the performance using our proposed ILBP-TOP outperforms that of approaches using LBP-TOP.

REFERENCES

[1] S. Heinrich and S. Wermter, Towards robust speech recognition for human-robot interaction, Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp.468-473, 2009.

[2] S. I. Yamamoto, K. Nakadai, M. Nakano, et al, Real-time robot audition system that recognizes simultaneous speech in the real world, Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 5333-5338, 2006.

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

[5] H. Li, J. Han, T. Zheng and G. 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.

[6] 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.

[7] 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.

[8] S. T. Shivappa, B. D. Rao and M. M. Trivedi, Audio-Visual Fusion and Tracking With Multilevel Iterative Decoding: Framework and Experimental Evaluation[J], IEEE Journal of Selected Topics in Signal Processing, vol. 4, no. 5, pp. 882-894, 2010.

[9] 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.

[10] 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

[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, pp. 22-23, 2004.

[12] P. S. Aleksic, J. J. Williams, Z. Wu and A. K. Katsaggelos, Audiovisual speech recognition using MPEG-4 compliant visual features, EURASIP Journal on Applied Signal Processing, vol. 11, pp. 1213-1227, 2002.

[13] 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, pp. 993-996, 2004.

[14] K. Saenko, K. Livescu, M. Siracusa, et al, Visual speech recognition with loosely synchronized feature streams, International Conference on Computer Vision, pp. 1424-1431, 2005.

[15] T. Ojala, M. Pietikainen and T. Maenpaa, Multiresolution gray scale and rotation invariant texture analysis with local binary patterns, IEEE Transactio on Pattern Analysis Machine Intelligence, vol. 24, no. 7, pp. 971-987, 2002.

[16] 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.

[17] Y. W. Wu, H. Liu and H. B. Zha, Modeling Facial Expression Space for Recognition, International Conference on Intelligent Robots and System, pp. 1814-1820, 2005.

[18] J. S. Lee, and C. H. Park, Adaptive decision fusion for audio-visual speech recognition, Speech recognition, technologies and applications, pp. 275-297, 2008.

[19] 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.

[20] M. Tariquzzaman, S. M. Gyu, K. J. Young, et al, Performance Improvement of Audio-Visual Speech Recognition with Optimal Reliability Fusion, International Conference on Internet Computing and Information Services, pp. 203-206, 2011.

[21] S. Tamura, K. Iwano, and S. Furui, A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization, IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 469-472, 2005.

[22] 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.

[23] 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.

[24] I. Matthews, J. A. Bangham and S. Cox, Audio-visual speech recognition using multiscale nonlinear image decomposition, Proceedings of the International Conference on Speech and Language Processing, pp. 38-41, 1996.

[25] T. W. Lewis and M. W. Powers, Sensor fusion weighting measures in audio-visual speech recognition, Proceedings of the 27th Australasian conference on Computer science. vol. 26, pp. 305-314, 2004.

[26] R. C. Rose, et al, A hidden Markov model based keyword recognition system, IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 129-132, 1990.

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

[28] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, 1989.

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

[30] G. Zhao, M. Barnard and M. Pietikainen, Lipreading with local spatialtemporal descriptors. IEEE Transactions on Multimedia, vol.11, no.7, pp. 1254-1265, 2009.