© 2017 IJSRSET | Volume 3 | Issue 5 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Keyword Spotting Based On Decision Fusion

M. Sowmya
Department of ECE, JNTUA College of Engineering , Ananthapuramu, Andhra Pradesh, India

ABSTRACT

Automatic speech recognition (ASR) technology is available now-a-days in all handsets where keyword spotting plays a vital role. Keyword spotting performance significantly degrades when applied to real-world environment due to background noise. As visual features are not affected much by noise this provides better solution. In this paper, audio-visual integration is proposed which combines audio features with the visual features where decision fusion used to adapt for various noise conditions. Visual features are extracted by a set of both geometry based features and appearance based features for facial landmark localization. To avoid similarities among the textons spatiotemporal lip feature (SPTLF) is used which map the features into intra class subspace. The dimensionality of the lip features are reduced using WPCA. A hybrid HMM-ANN method is proposed for integrating audio and visual features. Adaptive weights are generated using neural network for integration of audio and visual features. A parallel two step keyword spotting strategy is provided to avoid overlap between audio and visual keywords. Experiments results on dataset demonstrate that the proposed HMM-ANN method shows improved performance compared to the state of the art network.

Keywords: Automatic speech recognition (ASR), Keyword spotting, Decision fusion, WPCA, HMM-ANN method.

I. INTRODUCTION

Automatic speech recognition (ASR) deals with identification of words spoken by human. In recent decades, many researches have taken place. In continuous speech recognition, the complete transcription of input finds difficult, in such cases keyword spotting provides better compatibility. Keyword spotting deals with predefined keywords instead of complete utterance this also reduces the complexity of the system. There are three types of keyword spotting methods: LVCSR, acoustic and phonetic search KWS. Complete transcription of input is required in cases such as LVCSR. Contrast to LVCSR, acoustic based KWS method identifies only predefined keyword where key information lies only in some part of input utterance [2, 3]. Phonetic KWS converts speech to phoneme string and then calculate the distance for most appropriate keyword.

KWS finds applications in many areas such as audio data mining, home appliances activated by voice activated systems, visual passwords, speaker verification and as voice search engines. Although KWS technologies have achieved dramatic progress, most KWS systems are sensitive to noise such as background noises, machine noises and other voice activities when applied to real-world environment. Keyword spotting using both acoustic and visual information
is a solution to complementarily solve the problem.

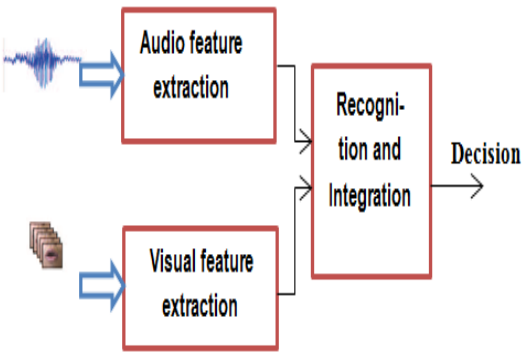

Figure 1: Basic block diagram of AV-ASR

II. REALTED WORK

The KWS performance is improved in many recent studies of which large vocabulary continuous speech

recognition (LVCSR) based KWS provides better performance as compared to acoustic KWS [2,3]. To adapt to various noise conditions an approach of rich system combination of KWS proposed in [2] where keyword spotting is done to withstand adverse and heterogeneous noise conditions. In adverse noise conditions the features of audio decreases dramatically, to overcome this we add visual features, which is susceptible to noise. In general communication with others, we not only listen to others speech but also look into the faces where visual feature also contributes. With the context of above scenario visual feature comes to existence, this type of features provide more information in cases such as hearing impaired or noisy situations. Many researches done in recent years of which recognizing the information by looking in to video without audio [3] provides better results.

The above figure 1 depicts basic block diagram of AV-ASR. Visual recognition provides information from the lip movements. In visual speech processing the video is the input where this is converted to frames and for each frame the face detection processing takes place. One of the approaches includes, the review of recent advances in visual speech processing in [3].VSR provides noise robustness as compared to ASR. Technique for lip reading [3] uses a simple model, which extracts visual features from video frames and plot in the curve to improve performance.

An English-oriented audio-visual KWS was proposed which adapts to noise [2]. Shivappa [4] proposed an intelligent meeting room for analysing the pose and information regarding the meeting. In [5] an integration technique used which includes PCA as feature extraction for mouth region and trained the data using the neural network. Many advances in audio-visual integration took place of which some concentrate on Mandarin language recognition and few on English language etc. In recent years, the audio-visual keyword spotting with the decision fusion was implemented [5], where the geometric features where used to localize the landmarks of the face such as eyes, nose and mouth.

The various methods for facial alignment are explained briefly below: (a) the geometric feature extraction depends on geometric properties such as height, width and length of lips, (b) in appearance based method, features are extracted based on pixel information of the mouth, (c) in image transform based methods the image is transformed into space using transformation techniques this majorly concentrates on edges, contours, corners etc. If local features are not much important appearance based approach is used as this considers lip features in global.

In visual feature extraction face by applying the appearance based model losses local information, which is overcome by disCLBP-TOP in earlier work. This provides the local information by using local binary pattern method. The dynamic information of mouth is mapped into three orthogonal planes. The mouth regions are converted to blocks where histograms of each block is computed and concatenated to single value. The sign and magnitude components are calculated after the feature extraction. In this the reliability model is computed by using the sigmoid function in order to adapt the noise conditions. This is followed by the integration of the audio and visual feature components. The main approach applied here is the disCLBP-TOP which is the extension of CLBP followed in earlier methods.

The paper contents are outlined as follows, section III. Section A describes the lip descriptor that is obtained by the shape difference feature and the spatiotemporal lip feature explained in section B. Optimal weights generation by neural network for audio-visual integration in section C. Results are included in Section IV followed by the conclusion V.

III. Proposed Work

In this paper audio-visual KWS is accomplished using neural network. In audio-based KWS MFCC are extracted from the acoustic features. The acoustic signal is divided into frames followed by the frequency analysis. In the frequency analysis each frame is segmented by hamming window of 25ms.Then the signal is passed to Mel filter bank followed by DCT transform. If any noise or signal distorts, it is removed by cepstral mean subtraction. The audio-based KWS is obtained from earlier works.

Visual speech recognition consists of following sections: First, the facial landmark is localized which is important for visual feature extraction. Second, shape difference features and spatiotemporal lip feature to capture textures and dynamics of lip movements process frames.

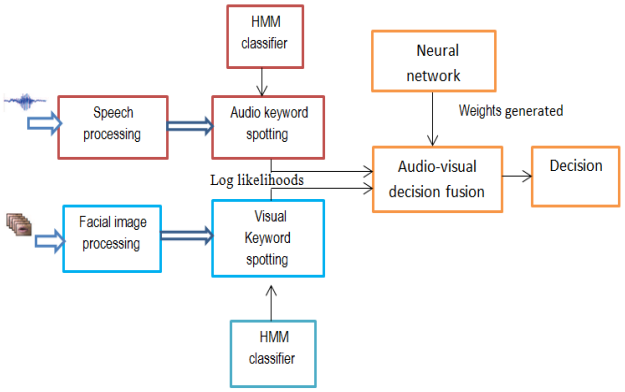

Figure 2: Parallel spotting strategy of AV-KWS

The below figure.3 depicts face detection from input video.

Figure 3: Facial Alignment

A. VISUAL FEATURE EXTRACTION

Facial Alignment

Facial landmark localization aligns facial parts, which is important for face recognition and face animation. Active appearance model is the conventional type which detects features based on pixel information. The limitation of this approach is highly sensitive to initialization due to gradient descent optimization. A regression based approach without using any parametric shape models is applied here. Assume that a face shape which consists of N facial landmarks. For given frame, facial landmark is estimated such that it utmost fits to the original landmark. First, boosted regression [11] is used to combine T weak regressor ($R^1, \ldots, R^t, \ldots, R^T$). Given a facial image I an initial face shape S^0, is assumed. Each regressor computes a shape increment ΔS from image features and then updates the face shape, which is formulated as

$$S = [x_1, y_1. \ldots x_N, y_N]^T \quad (1)$$

$$S^t = S^{t-1} + R^t(I, S^{t-1}),\ t=1,\ldots,T \quad (2)$$

Given N training samples, the regressor is learned until the training error no longer decreases. Each regressor R^t is learned as follows:

$$R^t = \arg\min_R \sum_{i=1}^{N} \|\hat{S}_i - (S_i^{t-1} + R(I_i, S_i^{t-1}))\| \quad (3)$$

here S_i^{t-1} is the previous estimated shape. The two-level cascaded regression and correlation-based feature selections [11] are used in this paper. In regression, face box is detected by bounding box followed by estimation of landmark in coarse-to-fine manner then the landmark is estimated as shown in Fig-3.

Then geometric centres of eyes and mouth are detected. Thus, lip region is cropped followed by feature extraction in next sub-section.

Shape Difference Feature

Lip region is cropped such that the geometry features such as width, height and contour of lips are evaluated in order to obtain the exact shape of the lips. The shape difference feature (SDF) provides the defined landmarks. Given M lip landmarks, four types of representations describe the lip shape here, as shown in Fig. 4 by calculating the Euclidean distance:

(a)Vector$\boldsymbol{d_1}$ provides the lip width and height (b)Vector$\boldsymbol{d_2}$ vertical distance between lips (c)Vector$\boldsymbol{d_3}$ provides the outer lip contour (d) Vector $\boldsymbol{d_4}$ inner lip contour distances

Shape difference feature vector $\boldsymbol{d}$ is computed as

$$d = [(\Delta d^1)^T, (\Delta d^2)^T, \ldots, (\Delta d^T)^T]^T$$

$$\Delta d^t = |d^{(t-1)} - d^{(t)}|,\ t=1,\ldots,T \quad (4)$$

where T is the number of frames from video.

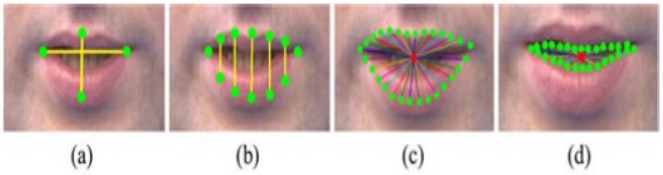

Figure 4: Information from mouth region (a) lip width and height (b) vertical distance (c) outer lip contour (d) inner lip contour

B. SPTLF

The utterance of similar word by different speakers varies. In order to overcome this we use spatiotemporal lip feature. This reduces the intra class variance and provides dynamic representation of lip feature of the following steps 1) Represent the cropped mouth region into blocks. 2) From these sample features extract low

level features 3) encode the features by locally constrained linear coding (LLC) 4) WPCA to reduce the dimensions. In this context we use textons defined as mini template of utterance with varying images in geometric configurations. DoG (Difference of Gaussian) used to enhance the low level features and eliminate noise in textons.

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 [10], a fast and effective coding strategy is used to encode the low-level patch-based features,

Image classification done by traditional SPM (spatial pyramid matching) has the drawback of poor scalability. In [9] vector quantisation (VQ) and sparse coding (SC) techniques are used. In this paper LLC [10] used which encodes the low level features using locality adaptor. Both LLC and SC represent descriptor by multiple bases such that similar descriptors have similar code whereas VQ uses single base. However, LLC can be performed very fast and requires less computational analysis compared to SC. B_k is the sub code book which represents the low level patch features of volume k. The codebook B is as follows:

$$B=B_k \mid \ k=1\ldots K \qquad (5)$$

$$B_k= [b_{k,1}, b_{k,2}, \ldots\ldots, b_{k,M}] \in R^{D\times M}$$

where M is the total number of entries in codebook, M>D. The low level patches are encoded by following criteria,

$$\min_c[\sum_{i=1}^{N} \| p_i - B_k\, ci\|^2 + \lambda \|d_i \Theta\ c_i\|^2]$$

$$\text{s.t. } 1^T c_i=1 \qquad (6)$$

where Θ indicates element multiplication. c_i is the reconstructed vector. Codes, C is the set of codes $[c_1,c_2,\ldots..,c_N]$ which assigns basis for each descriptor by

$$d_i=\exp\left(\frac{dist\left(p_i, B_k\right)}{\sigma}\right)$$

where, $dist(p_i,B_k)$ is the distance between p_i and $b_{k,j}$. σ is the weight adjustment determining the speed of the locality adaptor.

In coding K-means algorithm [10] is used to classify the clusters to define the similarity among the textons. In this cluster are of only two either keyword or non-keyword. First cluster, *k* values is initialised and Euclidean distance between centre of cluster and feature values are calculated and update the centre of cluster till it remains same. In order to obtain the statistical information of each volume, pooling is done. Here we use mix pooling technique to obtain the dynamic information of the textons which is trade-off between sum pooling and max pooling techniques.

$$x_k=\sum_{t=1}^{T} \max\, c_i \ \max_{c_i\in s_k^t} c_i \qquad (7)$$

Whitened Principal Component Analysis* (*WPCA):

The feature x_k is of high dimension this has to be reduced. Earlier works include PCA (principal component analysis) reduce feature by Eigen vectors corresponding to large Eigen values. The steps include: 1) map the features to intra class subspace by calculating the covariance matrix. 2) Λ be the Eigen vector of $[\lambda_1\ldots\ \lambda g]$ *g* is the large value and *V* be the corresponding Eigen vector. 3) The representation of y for volume k is given by

$$y_k=\text{diag}(\lambda_1^{-1/2},\ldots\ldots,\lambda_g^{-1/2})V^T x_k \qquad (8)$$

here responses from large Eigen values are supressed by multiplication with inverse of the feature. The decision feature for y for *k* volumes are given by

$$y=[y_1^T, y_2^T \ldots. y_k^T]^T \qquad (9)$$

here we apply this WPCA such that the feature d is represented by $f = [d^T, \mu.y^T]^T$ where μ is the adjustment factor which is obtained similar to. μ set to 1.2 to obtain better performance.

C. Integration of Audio And Visual Features

Integration

Integration of audio-visual features plays a vital role. Generally there are two fusion classifications: Feature fusion and Decision fusion. In feature level fusion both audio and visual features are combined first and then classified whereas in [5] decision level fusion each audio and visual features are classified first and then concatenated. Thus, earlier method requires large training data as single classifier employed with large dimension feature vector. In the later method each feature modelled explicitly and adapts to noise. In this

scheme late integration based decision fusion is employed.

Audio features are estimated by using hidden markov model (HMM) [5], where each voice data trained by using HMM model to obtain the features O_A.

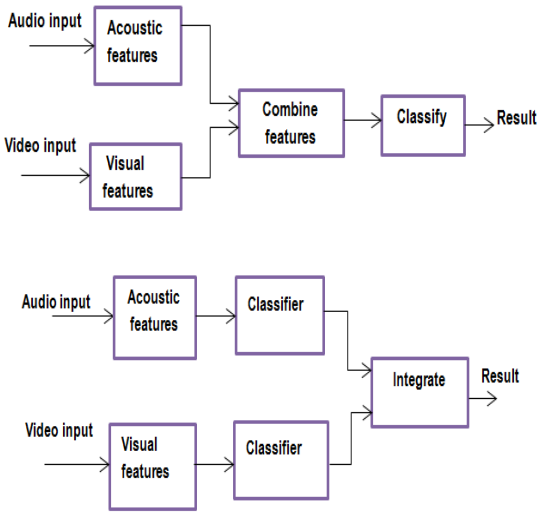

Figure 5: (a) Feature Fusion (b) Decision fusion

Similarly the visual feature trained and classified by HMM represented by O_V. Keyword spotting includes two steps to check whether keyword present in the speech in first step, followed by rejection of false alarms in second step. Thus, keywords which are not embedded in utterance gives results as zero whereas speech to text gives a garbage value. The neural network is used for integration, as synchronising the audio and visual features is difficult. Integration of audio and visual features by suitable weights generated by neural network is done as follows:

$$Log\ p\ (O_A/\ \lambda_{AV}) = \{\gamma\ log\ p\ (O_A/\ \lambda_i^A\ ) + (1-\gamma)\ log\ p\ (O_V/\ \lambda_i^V\ )\} \quad (10)$$

where γ is the adaptive weight generated by the neural network. O_A and O_V are the acoustic and visual feature of a keyword candidate while λ_i^A and λ_i^V are the acoustic and visual HMM of keyword .log $p(O_A|\lambda_i^A)$ and log $p(O_V|\lambda_i^V)$ represent acoustic and visual log-likelihood. γ is the function of audio and visual features to measure the reliability of the two modalities given as

$$\gamma = f(D_A, D_V) \quad (11)$$

where D_A and D_V are reliability of the two models. This is trained by feed forward neural network in such a way that optimal weight has to be obtained. This is trained for different resolutions of images followed by different noise conditions.

Synchronising of AV Features

A two-step keyword spotting is used in parallel to synchronise the audio and visual features. This should provide results in such a way that its performance should be equal to visual features if audio is more prone to noise and equal to audio features if SNR is low. The training steps are as follows: First, each of audio and visual features are obtained by using HMM model and are represented as vector. Second, based on integrating weights the audio and visual features are combined by neural network decision fusion. Third, based on the feature extraction keyword are recognised. The rejection rate of keyword is obtained from the difference log likelihood of keyword and filler.

If both audio and visual feature overlap in time they are directly removed as false alarm. If the middle point of one keyword modality falls in time with other keyword then it is said to be overlap in time. In those cases the keyword with large log likelihood are considered as true keyword else non keyword. If they don't overlap both are considered as keyword. In decision fusion neural network is used which consists of input, hidden and outer layers. In feed forward neural network if the hidden layers are unlimited provides faster and accurate recognition however this type has large time consuming and system complexity exists.

IV. RESULTS AND DISCUSSION

Acoustic feature Recognition:

In this speech is recorded at frequency of 16KHz. Mel frequency cepstral components are obtained and are trained by using HMM to obtain the OA features. The database includes words such as 'Information', 'Management', 'Arrangement', 'Moving', 'Grooving', 'Module',' 'Abacus', 'Forward', 'Backward', 'Identification'. In clean environment this is processed to obtain the result. With increasing SNR the recognition rate of audio increases. This provides less recognition in low SNR and vice versa. The accuracy comparisons are shown below in figure6.

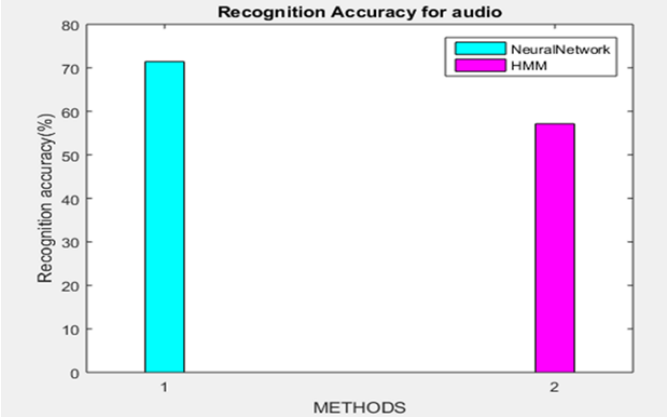

Figure 6: Accuracy for ASR –NN (neural network)

Visual Feature Recognition:

In this section 10 videos for database is recorded at 25 frames per second under normal lighting conditions. The same database is trained by HMM to obtain the features. First, the video is converted to frames. From each frame the face is detected using viola jones algorithm that employs Haar transforms to detect facial parts.

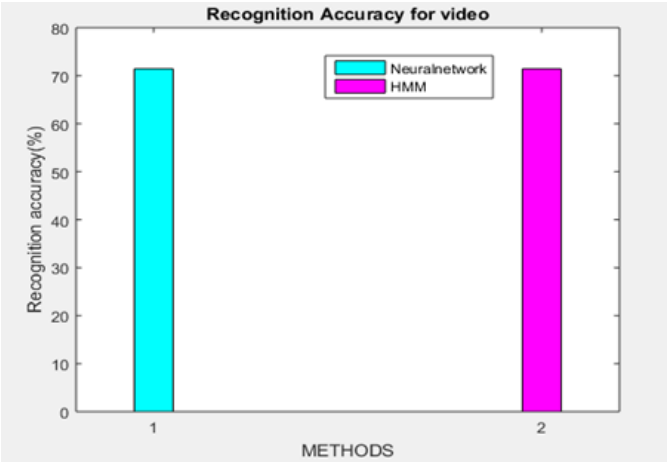

Figure 7: Accuracy for video HMM vs ANN

After which the alignment of facial landmarks takes place using shape regression analysis [11]. The accurate face alignment is required as this depicts the basic part for visual feature extraction. Second, from facial alignment mouth region is cropped and lip localisation algorithm is employed to obtain the centre of lips using Euclidean distance method. Third, decision fusion using neural network is employed.

Audio-Visual Integration

Only few databases are available for AV keyword spotting. Both the audio and visual features are trained by HMM filler model such that it provides better results even when no input is provided. FOM and recognition accuracy are used as parameters to evaluate the performance. FOM gives the ratio of correctly detected to the total input. In recognition accuracy (RA) defines the ratio of sum of TP and FP to the total.

RA=(TP+FP)/(TP+FP+TN+FN)

Audio-visual integration: Only few databases are available for AV keyword spotting. Both the audio and visual features are trained by HMM filler model such that it provides better results even when no input is provided. FOM and recognition accuracy are used as parameters to evaluate the performance. FOM gives the ratio of correctly detected to the total input. In Recognition accuracy (RA) defines the ratio of sum of TP and FP to the total.

RA= (TP+FP)/ (TP+FP+TN+FN)

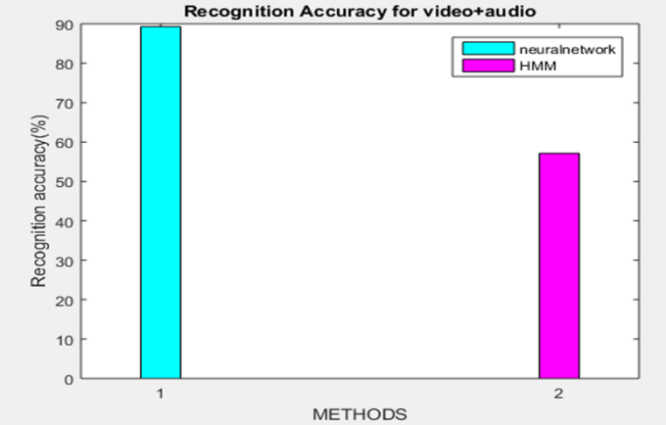

Figure 8: Accuracy for audio+video

For keyword spotting FP should be as low as possible. The neural network finds prior to HMM in cases where the computations required is less. The limitation of HMM is large data are required for training and also if new data is added retraining is required. In HMM, hidden states exist where as in neural network hidden computational units. In neural network RA parameter is calculated by using the formula given above.

The above figure-8 shows the accuracy for audio+video feature extraction comparing both the HMM and neural network. Experimental results validate that the codebook size of 512 in LLC provides maximum results as compared to the variations from 64, 128, 1024 In SPTLF the DoG filter parameter σ=2 with 8×4 patch size to achieve the better performance with mouth resolution of 100×180. The following table shows the

accuracy of different methods of KWS under different SNR conditions. The robustness to noise is evaluated by adding the noise in range of SNR 0dB, 5dB, 10dB, 15dB, 20dB. The accuracy for visual features remains same as shown in table-1.

SNR(dB)	0	10	20
Audio	15.51	30.52	74.52
Video	35.8	35.8	35.8
Audio+Video	40.08	63.61	89.28

Table-1: Comparisons of different methods for different SNR

V. CONCLUSION

The Hybrid HMM based keyword spotting method presented in this paper provides noise robustness when applied to real–world environment. In this paper the visual features extraction using SPTLF which provides the dynamic information of lips. Similar textons are coded by LLC using K-means thereby, reducing intra class variance. The large dimensions of lip feature are reduced by WPCA. The hybrid HMM-ANN techniques provides trade-off between HMM and ANN. The decision fusion based audio-visual integration adapts to noise as weights are generated accordingly using neural network. This finds application in HRI which gives better results even for untrained data using neural networks.

REFERENCES

[1]. A Novel Lip Descriptor for Audio-Visual Keyword based Spotting Based on Adaptive Decision fusion Ping Wu, Hong Liu, Member, IEEE, Xiaofei Li, Ting Fan, and Xuewu Zhang IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 18, NO. 3, MARCH 2016

[2]. P. Motlicek, F. Valente, and I. Szoke, "Improving acoustic based keyword spotting Using lvcsr lattices," in Proc. IEEE Int. Conf. Acoustic., Speech, Signal Process., Mar. 2012, pp. 4413–4416.

[3]. Z.Zhou, G. Zhao, and M. Pietikainen, "Towards a practical lip reading system," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., Jun. 2011, pp. 137–144.

[4]. S. T. Shivappa, B. D. Rao, and M. M. Trivedi, "Audio-visual fusion and tracking with multilevel iterative decoding: Framework and experimental evaluation," IEEE J. Sel. Topics Signal Process., vol. 4, no. 5, pp. 882–894, Oct. 2010.

[5]. J.-S. Lee and C.H. Park, "Adaptive decision fusion for audio-visual speech recognition," Speech Recog., Technol. Appl., pp. 275–296, 2008.

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

[7]. H.Liu, T. Fan, and P.Wu, "Audio-visual keyword spotting based on adaptive decision fusion under noisy conditions for human-robot interaction," in Proc. IEEE Int. Conf. Robot. Autom., May–Jun. 2014, pp. 6644–6651.

[8]. G. Zhao,M. Barnard, and M. Pietikai "Lip reading with local spatio temporal descriptors," IEEE Trans. Multimedia, vol. 11, no. 7, pp. 1254–1265, Jun. 2009.

[9]. J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., Jun. 2009.

[10]. J. Wang et al., "Locality-constrained linear coding for image classification," in Proc. IEEE Conf. Comput. Vis. Pattern Recog., Jun. 2010, pp. 3360–3367.

[11]. X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," Int. J. Comput. Vis., vol. 107, no. 2, pp. 177–190, 2014.