Middle-East Journal of Scientific Research 24 (2): 404-411, 2016
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2016.24.02.22850

Audio Video Recognition (AVR) Algorithm Based Labeling Features for Human Computer Interaction

[1]V. Anand and [2]R.S.D. Wahidha Banu

[1](Anna University) PSV College of Engineering and Technology,
Krishnagiri, Tamil Nadu, India
[2]Government College of Engineering, Salem, India

Abstract: In this paper proposed a Human computer Interaction based Audio video Recognition (AVR) algorithm designed to work in a realtime environment with multiple features classification approaches. Our work consist of a Color camera, Microphone and a depth sensor as input streams which is used to analyze the speech-related features approach consisting of Audio video Recognition (AVR) algorithm and Sound separation feature selection (SSS) modules. VSFS detects the features of face positioning history, eye gaze level and face annotation based face recognition. Currently, Audio video Recognition (AVR) based audio source separation has been shown to be a very strong algorithm in such scenarios, by dealing with the joint processing of two or more modalities of input data. The main idea is that by fusing multiple (audio and video) data streams it is possible to exploit the correlation among them in a way one modality compensates for the other's flaws. Based on these benefits, we propose a technique that uses three different input streams for developing an noisy and noiseless AVR algorithm, that runs real-time with an average 11.48% Error rate (DER) for scenarios presenting background noise, interfering sound sources and up to three simultaneous speakers.

Key words: Human behaviours · Man-Machine Interaction · Human computer Interaction · Upcoming computer

INTRODUCTION

This work deals with interaction design for a class of upcoming computer technologies for human use characterized by being different from traditional desktop computers in their physical appearance and the contexts in which they are used. Such technologies include for example wearable computers, context-aware computers, immersive virtual spaces and pervasive computerized environments and are typically referred to as emerging technologies. Emerging technologies often imply interaction dissimilar from how computers are usually operated. Consequently, such systems challenge the scope of established human-computer interaction styles and concepts and applicability of established methods and tools for their design.

The growth in Human-Computer Interaction (HCI) field has not only been in quality of interaction, it has also experienced different branching in its history. Instead of designing regular interfaces, the different research branches have had different focus on the concepts of multimodality rather than unimodality, intelligent adaptive interfaces rather than command/action based ones and finally active rather than passive interfaces.

As a result of the continuous advances in computing- related resources, Video and speech-based human-computer interfaces (HCI) have been receiving more attention from the scientific community lately, increasingly becoming more robust and practical for daily uses. Speech is a very promising modality for HCI, since it is in fact a very habitual (and thus efficient) communication mechanism used by humans, potentially being more effective than the current interfaces of common daily use (such as mouse, keyboard and touch screen), when applied successfully for HCI. Therefore, in order to implement speech-driven HCI that recognizes voice commands as accurately as humans, the task of automatic speech recognition (ASR) must be incorporated to the system.

Corresponding Author: V. Anand, (Anna University) PSV College of Engineering and Technology, Krishnagiri, Tamil Nadu, India.

Related Works: A number of algorithms and designs have been proposed in literature we shall discuss few of them here according to Human computer Interaction systems.

HCI design should consider many aspects of human behaviours and needs to be useful. The complexity of the degree of the involvement of a human in interaction with a machine is sometimes invisible compared to the simplicity of the interaction method itself. The existing interfaces differ in the degree of complexity both because of degree of functionality/usability and the financial and economical aspect of the machine in market. For instance, an electrical kettle need not to be sophisticated in interface since its only functionality is to heat the water and it would not be cost-effective to have an interface more than a thermostatic on and off switch. On the other hand, a simple website that may be limited in functionality should be complex enough in usability to attract and keep customers [1].

Therefore, in design of HCI, the degree of activity that involves a user with a machine should be thoroughly thought. The user activity has three different levels: physical, cognitive [2] and affective [3]. The physical aspect determines the mechanics of interaction between human and computer while the cognitive aspect deals with ways that users can understand the system and interact with it. The affective aspect is a more recent issue and it tries not only to make the interaction a pleasurable experience for the user but also to affect the user in a way that make user continue to use the machine by changing attitudes and emotions toward the user [4-6].

Sometimes called as Man-Machine Interaction or Interfacing, concept of Human-Computer Interaction/Interfacing (HCI) was automatically represented with the emerging of computer, or more generally machine, itself. The reason, in fact, is clear: most sophisticated machines are worthless unless they can be used properly by men. This basic argument simply presents the main terms that should be considered in the design of HCI: functionality and usability [7].

Typical existing SD approaches deal with the "who spoke when?" problem, as previously mentioned. This implies the off-line processing of formerly recorded streams and is normally applied for the purposes of automatic document annotation, as in speech transcription, speaker indexing and speech recognition. Existing works aiming towards such applications generally categorize as the so called *Bottom-up* approaches. These are based on agglomerative hierarchical clustering (AHC) techniques, where the clusters are commonly modeled as Gaussian mixture models (GMM) having short-time mel-frequency cepstral coefficients (MFCCs) as features and using a stopping criteria such as the Bayesian Information Criterion (BIC) or Kullback-Leibler (KL)-based metrics.

Why a system is actually designed can ultimately be defined by what the system can do i.e. how the functions of a system can help towards the achievement of the purpose of the system. *Functionality* of a system is defined by the set of actions or services that it provides to its users. However, the value of functionality is visible only when it becomes possible to be efficiently utilized by the user [2]. *Usability* of a system with a certain functionality is the range and degree by which the system can be used efficiently and adequately to accomplish certain goals for certain users. The actual effectiveness of a system is achieved when there is a proper balance between the functionality and usability of a system [3].

In [5] on-line SD is approached as a combination of off-line and on-line algorithms. The off-line system is a bottom-up technique based on [8, 9] and is used to cluster different speaker regions by processing all available data up to the current time. The on-line system then uses a Maximum a Posteriori technique to adapt a pre-trained set of generic speaker models to the clustered regions (namely, universal background models (UBM) [7]). MFCC features are extracted from the audio stream and majority voting is used for assigning short-time audio segments to a speaker model. The authors present good results in terms of accuracy, but the off-line module requires a one minute preparation and the on-line decision is performed with 2.6 seconds latency (real-time for their system).

Markov and Nakamura proposed an on-line SD system in [10], later improved in, that is capable of adapting to the acoustic scenario by recurrently managing the speaker models. A set of GMMs with the likelihood ratio test is first applied for VAD, at which point gender identification is also performed. If speech is detected, another set of GMMs is used to identify which speaker the audio segment belongs to. In case a new speaker is identified, a variant of the Expectation-Maximization (EM) algorithm is used to train a GMM for the new speaker using a long-term segment of the audio data. For all the classification modules, MFCC features are used and models unused within a certain time-window are discarded. While the authors claim the results are comparable to those of the state-of the- art, their system runs on latency between 3 and 4 seconds (real-time) and do not deal with overlapping speech [11-18].

In Noulas and Krose developed a multimodal on-line SD system for HCI purposes. Potential speakers are found using a face detector and then their audio-visual behavior is modeled as states of a dynamic Bayesian network. For the observations, Scale-invariant feature transform (SIFT) is applied to the facial features extracted from the images and the MFCC features are extracted from the audio. The models of each speaker are updated through hierarchical model selection [8] as more data arrive. The authors claim satisfactory results, but the on-line processing is still slower than real-time and it does not deal with overlapping speech. Furthermore, experiments were limited to two short test scenarios.

Another on-line multimodal SD system is proposed by Hung and Friedland [12], focusing on meeting scenarios with four speakers. Taking a previous recording as reference, they use the work in [17] (an off-line system) to create and train a set of GMMs for each speaker and one for non-speech situations. These models are then used to process the remaining recordings, where speech is assigned to the users by a likelihood test using cepstral coefficients as audio features. Video is later used for labeling the speakers from body movement information that is obtained from the residual coding bit-rate of the video compression algorithm. The speakers' audio-based models are associated with the visual features from correlation measures between the modalities. This approach runs on 2.2 second latency and presents a low DER, but assumes a known number of participants and also does not deal with overlapping speech.

In a real-time SD technique is proposed. The authors modify the system in which is a bottom-up approach of the category previously described. Two techniques are presented which are able to reduce the cluster merging hypothesis space by up to 86% before advancing to the BIC merging step, which is considered the most expensive part of such typical SD systems. One technique is based on pitch-correlogram [19] and the other on KL-divergence. Clusters deemed unlikely to merge are inexpensevely excluded in such step. Although the system runs in real-time keeping its original accuracy, it does not process the data on-line (meaning the entire audio stream is analyzed).

From the works mentioned so far, we notice that on-line SD is still a challenge where trade-offs must be made to achieve reasonable results. Most of them primarily sacrifice latency for diarization accuracy, explaining why off-line systems, in general, present better results than on-line ones [3, 5]. Other drawbacks are usually related to the number of participating speakers. If it is assumed to be known, a model adaptation step before the system's operation is necessary. In particular, this is common in on-line SD, as in the works of Kwon and Narayanan, where the generic adaptable speaker model approach (namely UBMs) is proposed and later used to develop unsupervised SD methods [14], however at the cost of requiring an initialization step. While the UBM approach is promising in some cases, it is infeasible for short recordings where few adaptation data is present. In addition, these systems do not address the problem of overlapping/simultaneous speakers. This is very important in realistic scenarios where there can exist multiple users simultaneously providing speech inputs to the system (e.g., gaming), or also when non-users accidentally activate the system by speaking nearby.

Proposed Avr Methodology: Given the HCI focus in our work, in which on-line and realtime processing are required, the mentioned typical SD problems must be overcome. We have opted for a partially supervised approach, where Audio Video Recognition (AVR) is performed by an SVM classifier and SSS and speaker labeling by unsupervised ones. The set of features used for Audio Video Recognition (AVR) are also distinct from the conventional MFCCs. We extract acoustic, spatial and visual information from multiple data streams (an 8-element microphone array, a color camera and a depth sensor) to represent voice activity in multiple user scenarios with overlapping speech. SSS is performed by a 3D face-tracking algorithm based on [50] and a hybrid approach is applied for identifying/labeling the speakers.

Fig. 1 shows a diagram of our algorithm's execution flow, where the green boxes indicate the steps where SD is obtained as a natural consequence after they are concluded. For the capture system, we employ an eight-element linear microphone array as the main audio stream, along with a Kinect device for the RGB and depth streams. The RGB camera is aligned with the array's center and is considered to be the origin of our coordinate system. Our setup expects the users to be facing the capture sensors and to be within the camera's field of view as is the case of most HCI systems.

Fig. 2 illustrates our setup in more details. Another assumption in our system is that at least two microphones are available, so that SSS techniques may be applied.

For detailing the algorithmic part of our combined AUDIO VIDEO RECOGNITION (AVR) and SSS, the following division is used in the next subsections: (a) video feature extraction, (b) audio feature extraction, using AUDIO VIDEO RECOGNITION (AVR).

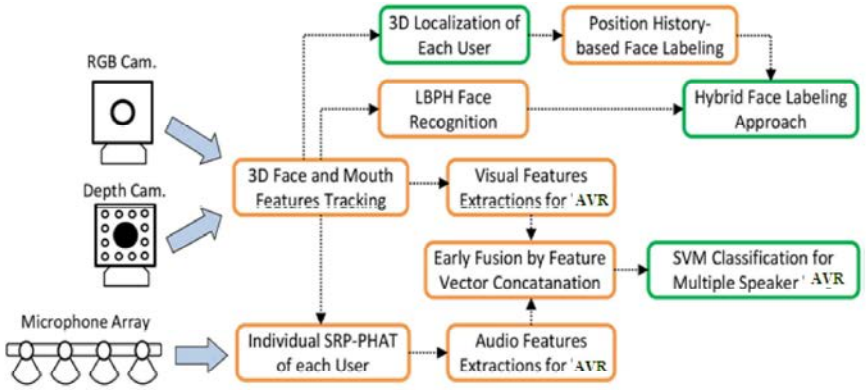

Fig. 1: General Architecture Of Proposed AVR Design.

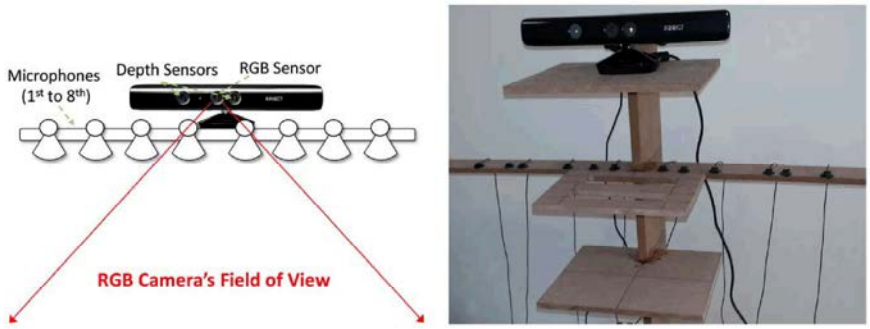

Fig. 2. (a) Schematic representation of the proposed approach. (b) Our prototype system.

Video Feature Extraction: To extract reliable features from visual information we exploit the fact that people with intention to speak move their lips. Assessing this information has been previously performed in the literature in some different ways: by computing the optical flow of mouth regions, measuring mouth geometries based on color information, or by detecting the lip's contours through active shape models (ASM). In our method, we opt for the ASM approach given that it has shown, in comparison to our previous work, to be the most robust technique for dealing with head rotation and face scale changes. The method we use is available in Microsoft's latest Kinect Face Tracking SDK, which implements the face model to express the facial features.

By running the algorithm on both color and depth frames as inputs, it is possible to initially identify number of users within the camera's field of view (FOV), who correspond to the potential speakers in the scene. Furthermore, a set of facial features for each person is returned by the tracker, which amounts to a total of 113 3D points spanning the ear to ear contour, eyes, eyebrows, inner lips, lower lips and nose. Consequently the 3D location, for, of each user is also retrieved from such points, where stands for "Kinect". For AUDIO VIDEO RECOGNITION (AVR) purposes, the subset of points that are most discriminative are those covering the lips' contour, intuitively. Thus, they are the ones from which we chose to derive the visual features. In Fig. 3 we may observe an example frame of our algorithm's execution, which displays the referred detected mouth regions.

From the figure, we may observe that both the inner and outer lip contours are identified. By experimenting with these two possibilities, we found the inner lips to be more representative for AUDIO VIDEO RECOGNITION (AVR) purposes and that by using both no extra precision is achieved (only redundant data is added). We rationalize this to the fact that when speaking, vertical mouth movements tend to stretch the inner contour more than the outer one. While normal open/close movements (which are predominant) are well represented by either contours, it is clearly beneficial to account for as many movement patterns as possible. Therefore, for each of the users, we extract a total of four features from the inner lips contours, which are described next.

By denoting and as the $d_w(t)$ width and $d_h(t)$ height of a given users' mouth, respectively in the image at the th frame, the features used for the AUDIO VIDEO RECOGNITION (AVR) classifier are computed as follows

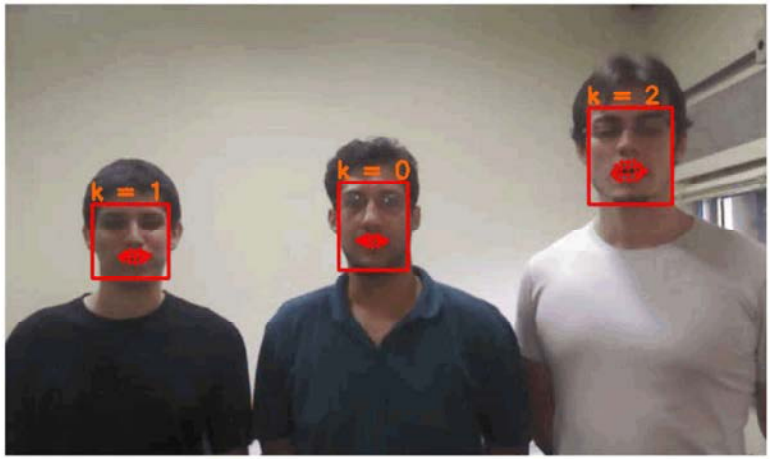

Fig. 3: Output example of the face tracker AVR algorithm. Faces' bounding boxes and lip contours are used in our approach.

$$g_1^{\mathrm{vid}}(t) = \frac{1}{T}\sum_{i=0}^{T-1} d_{\mathrm{w}}(t-i)$$

$$g_2^{\mathrm{vid}}(t) = \sqrt{\frac{1}{T-1}\sum_{i=0}^{T-1}\left(d_{\mathrm{w}}(t-i) - g_1^{\mathrm{vid}}(t)\right)^2}$$

$$g_3^{\mathrm{vid}}(t) = \frac{1}{T}\sum_{i=0}^{T-1} d_{\mathrm{h}}(t-i)$$

$$g_4^{\mathrm{vid}}(t) = \sqrt{\frac{1}{T-1}\sum_{i=0}^{T-1}\left(d_{\mathrm{h}}(t-i) - g_3^{\mathrm{vid}}(t)\right)^2}.$$

Although using this temporal evaluation process increases the robustness of the video features to the described cases, they may still not be enough for the cases when the visual modality becomes unreliable. One case would be scenes with distant capture, where facial expressions become unidentifiable to the point they are poorly tracked. Another instance would be people that move (translate or rotate) their head constantly, causing inaccurate tracking of the lips and thus undesired mouth movement representation by the features. In occasional situations, some people may also present very little mouth movement when speaking, degrading the features too. Therefore, the audio modality is the key for compensating such adversities for making the final AUDIO VIDEO RECOGNITION (AVR) algorithm more robust.

Audio Feature Extraction: The problem of voice activity detection has been predominantly approached through audio analysis in the literature. This is innate given the presence of speech can only be truly verified by processing sound (again, mouth movements alone may not produce / represent speech, meaning the video modality could easily be tricked).Even though such classes of methods may present robustness to noisy acoustic scenarios, they tend to fail in multiple speaker cases, where competing sound signals overlap in the time-frequency domain. In other words, the voice from a speaker acts as noise that has a very similar spectral pattern to the voice of other speakers. Furthermore, some of the mentioned works are not derivable from the single to multiple speakers case, given the use of a single microphone at close range for voice capture.Therefore, the common choice to enhance the capabilities of an audio-based classifier is to employ a microphone array as the audio stream, which allows the use of beamforming techniques to filter environmental noise and better isolate different speakers we derive our audio features from computations of the SRP-PHAT database, which is an Sound source separation (SSS) technique well known for its robustness under high noise and reverberation. The method essentially finds the point in space that produces the highest PHAT-weighted acoustic power. More specifically, for a frame of data, a search region is determined and the output powers of PHAT-weighted filter-and-sum beamformers, steered towards each candidate location, are compared; the location containing the highest SRP-PHAT value represents where the dominant sound source is positioned. Mathematically, this is defined by the following equation:

$$\hat{\mathbf{q}} = \arg\max_{\mathbf{q}\in\mathcal{Q}} P(\mathbf{q})$$

$$P(\mathbf{q}) = \int_0^{2\pi} \left| \sum_{m=1}^{M} \frac{X_m(\omega)}{|X_m(\omega)|} e^{-\mathrm{j}\omega\tau_m^{\mathbf{q}}} \right|^2 d\omega.$$

The robustness of the method owes mostly to the PHAT weighting function, which is represented by the denominator term. It equally emphasizes all frequency components of the sound signal by normalizing their amplitudes.

$$g_1^{\text{aud}}(t) = \frac{1}{T}\sum_{i=0}^{T-1} P(\mathbf{q}(t-i)_k^{\text{srp}})$$

$$g_2^{\text{aud}}(t) = \sqrt{\frac{1}{T-1}\sum_{i=0}^{T-1} P(\mathbf{q}(t)_k^{\text{srp}}) - g_1^{\text{aud}}(t))^2}$$

$$g_3^{\text{aud}}(t) = \frac{1}{T}\sum_{i=0}^{T-1} d_{\text{loc}}(t-i)$$

$$g_4^{\text{aud}}(t) = \sqrt{\frac{1}{T-1}\sum_{i=0}^{T-1} (d_{\text{loc}}(t) - g_3^{\text{aud}}(t))^2}$$

Audio Video Features Based - Avr Labeling: At this point, the Audio Video Recognition (AVR) and SSS modules of our approach have been described. Alone, they are enough to jointly perform multi-speaker Audio Video Recognition (AVR) and SSS, sufficing the described problem in Section III (sound events are validated through Audio Video Recognition (AVR) and users are distinguished through SSS). However, as is, the system would only be capable of associating speech segments to the participants based on their location, instead of their identity within the application. In other words, a speech segment would be said to belong to "the user at location …", while it should be "to the user with label…". This may cause problems, for example, in the situation where people move during the recording. Such people would receive multiple (possibly repeated) labels based on the different occupied locations. This scenario does not characterize a diarization approach, implying it is inadequate fortasks such speech annotation/indexing, or automatic user profiling. The latter is rather noticeable, e.g., in applications where the system continuously adapts to specific participants based on current user experiences. In order to achieve this functionality, a speaker labeling (SL) approach must be incorporated to the existing Audio Video Recognition (AVR)/SSS technique. Such approach must be able to keep track of speakers' labels throughout the entire system run (which in our case, is a multimodal recording), meaning that participants should not have their labels changed and new users who join the scene in the middle of the recording should receive new labels. It is important to notice that by label we do not necessarily mean the identity of the person (as its name), but simply some tag that is unique and persists to each person during the system's lifetime.

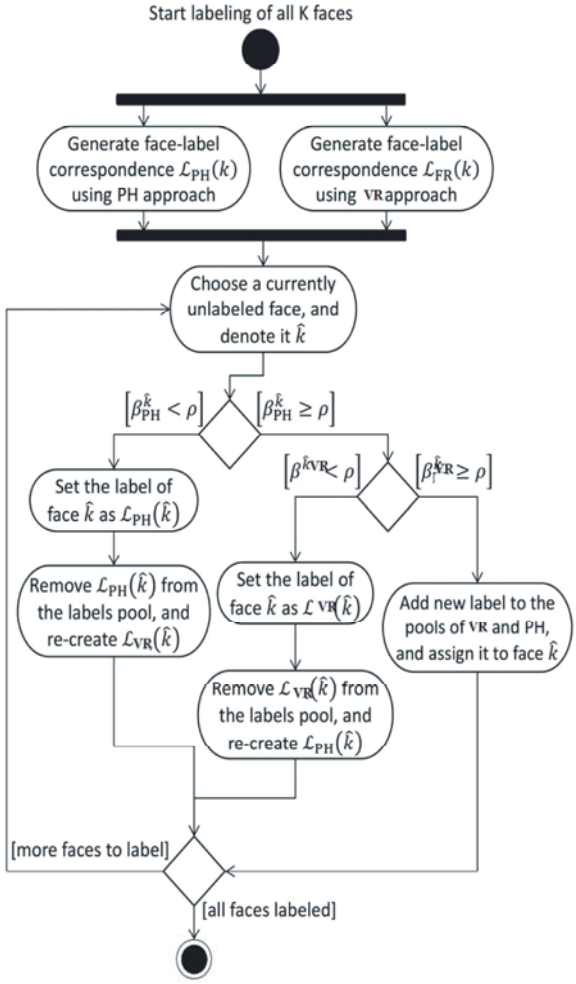

Fig. 4: Activity diagram of our AVR labeling approach

To achieve identity recognition, a database relating face Fig. 4. Activity diagram of our labeling approach. samples to names would be required, which is too specific to the system's set of users and would make the technique highly dependent on the environment. The proposed SL approach combines a Voice recognition (VR) method and the position history (PH) of the users to achieve the described demand. First, the PH approach is used to assign a label to each speaker and later the VR is used to refine the results and also detect unprecedented users (who should be getting new labels). Fig. 4 shows the flowchart of our labeling module. The variables in the diagram are explained next.

The Position History- AVR Based Approach: The PH approach starts by storing the 3D location of each face in an individual circular array of size, such that the most recent locations of each user are available at each time frame. Consequently, there will always be or more circular

arrays within the application's lifetime, meaning if a face is lost, will decrease, but the number of circular arrays will not (allowing users to re-join or new users to enter the scene). Let us denote such circular arrays as, for, where is the maximum number of detected faces up to the present frame (recalling that is the number of faces at the current frame) and the superscript "PH" denotes "position history". At each new frame, an average of the 3D positions is computed from each circular array through

$$\mathbf{q}_n^{\mathrm{PH}} = \frac{1}{U}\sum_{u=1}^{U}\mathbf{Q}_n^{\mathrm{PH}}(u)$$

The Voice Recognition- (AVR & SSS)Based Approach: Despite such cases being uncommon, we still handle them by validating the distance-based matching through an VR method. In this work, we used the LBPH-based, which is particularly favorable in comparison to traditional others because it allows the trained face models to be incrementally updated, being appropriate for an online sound source seperation(SSS) approach (does not cause dependency on a pre-computed face database). Essentially, this VR constructs a concatenated feature histogram for each face label by dividing face images into sub-regions and extracting circular local binary patterns as features. The update procedure simply works by creating more histograms from a given face's sub-regions and the recognition process is performed through a nearest neighbor classifier with chi-square as a dissimilarity measure.

$$b_{k,n} = \min_{v=1,\ldots,V_n} \chi^2\left(H_k^{\mathrm{kin}}, H_n^v\right)$$

$$\mathcal{L}_{\mathrm{final}}(k) = \begin{cases} \mathcal{L}_{\mathrm{PH}}(k), & \beta_{\mathrm{PH}}^k < \rho \\ \mathcal{L}_{\mathrm{VR}}(k), & \beta_{\mathrm{VR}}^k < \rho \le \beta_{\mathrm{PH}}^k \\ N+1, & \text{otherwise.} \end{cases}$$

RESULTS AND DISCUSSION

We analyze the accuracy of our proposed speaker labeling approach. In a similar fashion as done for Audio Video Recognition (AVR), the individual PH and VR approaches are compared to the hybrid one. For such experiments, recordings with only one speaker are not included, for the labeling is trivial. Table I displays the results, from which we may notice the proposed approach is close to 100% accurate. While both PH and VR perform reasonably well individually, fusing them through the scheme of Fig. 4 produces a superior labeling algorithm, such that the adverse situations previously mentioned for

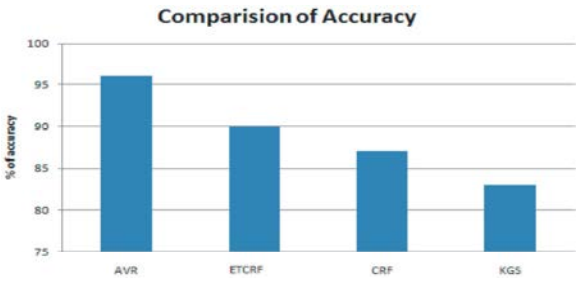

Fig. 5: Performance Comparison of proposed HCI

Table 1: AVR accuracy of our proposed approach, compared to ones Using only audio and video features for t SSS features set

Sequence	Hybrid	PH	VR
Two1	100.00%	100.00%	97.98%
Two2	100.00%	98.91%	99.54%
Two3	100.00%	100.00%	96.11%
Two4	100.00%	99.40%	97.54%
Two5	99.10%	98.56%	97.27%
Two6	100.00%	99.58%	96.42%
Two9	100.00%	98.28%	99.01%
Two10	100.00%	100.00%	97.07%
Two11	100.00%	99.12%	97.12%
Two12	100.00%	98.77%	89.46%
Two13	100.00%	98.98%	90.49%
Two14	100.00%	99.02%	91.91%
Three1	99.46%	97.41%	97.54%
Three2	99.10%	98.66%	98.84%
Three3	100.00%	99.54%	98.03%
Three6	100.00%	99.45%	98.33%
Three7	99.71%	99.02%	97.44%
Average	99.85%	98.23%	96.48%

SL are overcome. This is rather apparent for and, where VR fails due due to the several partial face occlusions, but is improved by PH in the hybrid approach. Altogether, having a precise labeling algorithm is very important so that the correct results of the Audio Video Recognition (AVR) one are not misused, that is, speech segments properly detected do not get assigned to the wrong speakers.

As a final set of experiments, the effectiveness of our SD algorithm is assessed in terms of DER. Such measure comprises both Audio Video Recognition (AVR) and labeling, allowing SD methods to be evaluated as a whole. It corresponds to the fraction of time that is not attributed correctly to one or more speakers (or to none, in case of silence), that is, the composition of the following measurements: (a) false alarm error, when speech has been incorrectly detected by the system; (b) miss error, when a person is speaking but the system fails to detect his/her voice activity; (c) speaker labeling error, when a given speaker receives a label not matched by the ground truth mapping.

From the comparison results shows that the Proposed AVR has better evaluation of ability of each feature in distinguishing of video and Audio classes with the accuracy of 96%.

CONCLUSION

We targeted HCI situations, where a response needs to be promptly presented to the users, (details of the concerned scenario was made available online as a link to our dataset). While our method deals with up to three simultaneous speakers, future work will aim on testing it in a more challenging scenario, with four (or more) overlapping speakers. Reducing the amount of microphones and generalizing the AVR, SSS model (using a lager data set for training) are other challenging goals that would help make the system a practically deployable commercial product. Finally, experimenting with different AVR features (such as Video or short-term audio energy) is also included in our subsequent work.

REFERENCES

1. Jaimes, A. and N. Sebe, 2007. Multimodal human-computer interaction: A survey,Comput. Vis. Image Understand., 108(1-2): 116-134.
2. Rabiner, L. and B.H. Juang, 1993. Fundamentals of Speech Recognition, Upper Saddle River, NJ, USA: Prentice-Hall.
3. Anguera Miro, X., 2012. Speaker diarization: A review of recent research, IEEE Trans. Audio, Speech, Language Process., 20(2): 356-370.
4. Brandstein, M. and D. Ward, 2001. Microphone Arrays: Signal Processing Techniques and Applications, ser. Digital Signal Processing. New York, NY, USA: Springer.
5. Moattar, M. and M. Homayounpour, 2012. A reviews on speaker diarization systems and approaches, Speech Commun., 54(10): 1065-1103.
6. Barras, C., X. Zhu, S. Meignier and J. Gauvain, 2006. Multistage speaker diarization of broadcast news, IEEE Trans. Audio, Speech, Language Process., 14(5): 1505-1512.
7. Vijayasenan, D., F. Valente and H. Bourlard, 2009. An information theoretic approach to speaker diarization of meeting data,IEEE Trans. Audio, Speech, Language Process., 17(7):1382-1393.
8. Asoh, H., 2004. An application of a particle filters to Bayesian multiple sound sources tracking with audio and video information fusion, in Proc. Int. Conf. Inf. Fusion, pp: 805-812.
9. Butko, T., A. Temko, C. Nadeu and C. Canton-Ferrer, 2008. Fusion of audio and video modalities for detection of acoustic events, in Proc. Interspeech, pp: 123-126.
10. Almajai, I. and B. Milner, 2008. Using audio-visual features for robust voice activity detection in clean and noisy speech, in Proc. 16th Eur. Signal Process. Conf., pp: 1-5.
11. Petsatodis, T., A. Pnevmatikakis and C. Boukis, 2009. Voice activity detection using audio-visual information, in Proc. 16th Int. Conf. Digital Signal Process., pp: 1-5.
12. Friedland, G., H. Hung and C. Yeo, 2009. Multi-modal speaker diarization of real-world meetings using compressed-domain video features, in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., pp: 4069-4072.
13. Schmalenstroeer, J. and R. Haeb-Umbach, 2010. Online diarization of streaming audio-visual data for smart environments, IEEE J. Sel. Topics Signal Process., 4(5): 845-856.
14. Schmalenstroeer, J., M. Kelling, V. Leutnant and R. Haeb-Umbach, 2009. Fusing audio and video information for online speaker diarization, in Proc. Interspeech, pp: 1163-1166.
15. Noulas, A., G. Englebienne and B. Krose, 2012. Multimodal speaker diarization, IEEE Trans. Pattern Anal. Mach. Intell., 34(1): 79-93.
16. Vallet, F., S. Essid and J. Carrive, 2013. A multimodal approach to speaker diarization on TV talk-shows, IEEE Trans. Multimedia, 15(3): 509-520.
17. Rouvier, M., G. Dupuy, P. Gay, E. Khoury, T. Merlin and S. Meignier, 2013. An open-source state-of-the-art toolbox for broadcast news idolization, in Proc. Interspeech, pp: 1477-1481.
18. Vijayasenan, D. and F. Valente, 2012. Diartk: An open source toolkit for research in multistream speaker diarization and its application to meetings recordings, in Proc. Interspeech, pp: 2170-2173.