Development of Framework for Automatic Speech Recognition

Prof. S.Qamar Abbas
Ambalika Institute of Management & Technology,
Lucknow,Uttar Pradesh, India,
qrat_abbas@yahoo.com

Nidhi Srivastava
Amity Institute of Information Technology,
Amity University
Lucknow, Uttar Pradesh, India
nidhinavya@gmail.com

Abstract:
In this paper we have proposed an automatic speech recognition framework using agents. In this we have included both audio recognition and visual recognition. The audio and visual modalities are complementary to each other and the combination of the two can improve the accuracy in affective user models. The audio features extracted are processed by audition agent. The visual processing agent takes care of the lip and face detection. Finally both these agents assist audio visual fusion agent in fusion of these modalities for automatic speech recognition.

Keywords: Agents; Audio-visual; Speech recognition; Face detection; Lip motion; Framework

I. INTRODUCTION

Vision plays a major role in human-computer interfaces that aim at natural interaction. Visual information from the speaker's mouth speech has been successfully shown to improve noise robustness of automatic speech recognizers, thus promising to extend their usability in the human computer interface. Automatic speech recognition is viewed as an integral part of future human-computer interfaces, which are envisioned to use speech, among other means, to achieve natural, pervasive and ubiquitous computing [1].

The visual channel carrying facial expressions and the auditory channel carrying vocal intonations are widely thought of as most important in the human recognition of affective feedback. According to Mehrabian, whether the listener feels liked or disliked depends only for 7% on the spoken word, for 38% on vocal utterances, and for even 55% on facial expressions. This indicates that, while judging someone's affective state, people rely less on body gestures and physiological reactions displayed by the observed person; they rely mainly on his facial expressions and vocal intonations. As far as body gestures are concerned, as much as 90% of body gestures are associated exclusively with speech. Hence, it seems that they play a secondary role in the human recognition of affective states [2].

Audio is probably the most natural modality to identify a speaker. However, video also contains important biometric information, which includes still frames of face and temporal lip motion information that is correlated with the audio. Most speaker identification systems rely on audio-only data. However, especially under noisy conditions, such systems are far from being perfect for high security applications. The same observation is also valid for systems using only visual data; where poor picture quality, changes in pose and lighting conditions or varying facial expressions may significantly degrade performance. Hence a robust and precise solution should employ all available sources of information in a unified scheme [3].

II. PROPOSED FRAMEWORK

The agent is a new concept and technology in the field of software development that can substitute user for producing some intelligent behavior [4].

The agent is provided with autonomous synergic and learning properties. Autonomous property means that agent may exercise control over its own actions, and do not need people's assistance. Synergic collaborate for accomplishing collective task learning property means that agent can achieve knowledge from interaction with people or other agent. Software agent is independent entity circulated in particular environment and many agents can collaborate for accomplishing collective task. Therefore, software agent especially adapt to human computer interaction. Because, user can exchange information with computer by using audio, video and lip motion used for various perception methods must possess some cognitive function [6**].**

Various agent forms the base for multimodal interaction interface and agents can circulate on a computer.

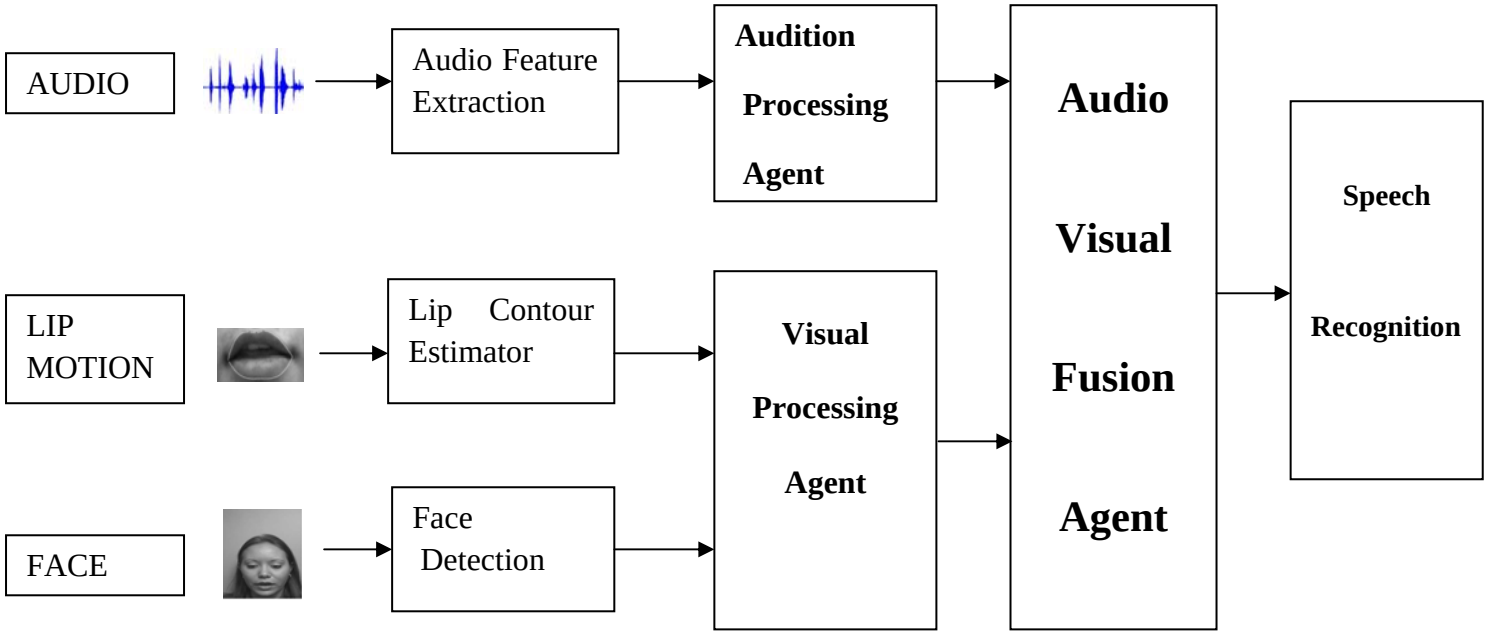

Fig. 1: Automatic Speech Recognition Framework

2.1 AUDIO FEATURE EXTRACTION

Speech is a necessary modality to enable a pervasive and consistent user interaction with computers across different devices --- large or small, fixed or mobile, and it has the potential to provide a natural user interaction model. However, the ambiguity of spoken language, the memory burden of using speech as output modality on the user, and the limitations of current speech technology have prevented speech from becoming the choice of mainstream interface [11]. Speech content and voice can be interpreted as two different, though correlated, information existing in audio signals [7]. The emotional information in the voice depends on the subject and recording condition. The pitch varies widely from person to person. In general, males speak with a lower pitch than females. Thus, for a given subject, the pitch at every frame is normalized by the pitch mean of the neutral expression sequence of the same subject [12].

The auditory features usually estimated from the input audio signal are:

1) *pitch* (the fundamental frequency of the acoustic signal delimited by the rate at which vocal cords vibrate);

2) *intensity* (the vocal energy);

3) *speech rate* (the number of words spoken in a time interval; words can be identified from time-varying spectra of harmonics, which are generated by vocal cord vibrations and filtered as they pass through the mouth and nose);
4) *pitch contour* (pitch variations described in terms of geometric patterns);
5) *phonetic features* (features that deal with the types of sounds involved in speech, such as vowels and consonants and their pronunciation) [2].

2.2 FACE DETECTION

Humans make use of face as an important cue for identifying people. Face detection and facial expression is very important in multimodal human computer interaction. The automatic recognition of natural facial expressions is a necessary step [8]. Facial expression conveys both verbal and non-verbal information. Face detection is a challenging task since there are many conditions that may vary. Each person has a unique face, meaning that each face looks different. Even the face of the same person looks different depending on the time when the image is taken.eg. the age of the person, eyeglasses, beard, moustache and make-up make a difference [9].

2.3 LIP CONTOUR ESTIMATOR

Lip information has been extensively employed in the state-of-the-art audio-visual speech and speaker recognition applications, since lip movements are highly correlated with the audio signal. Speech content can be revealed through lip reading; and lip movement patterns also contain information about the identity of the speaker.
The accuracy and robustness of the lip contour extraction method are crucial for a recognition system that uses lip shape information. Many techniques are available that attempt to solve the lip segmentation/tracking problem. The performance of these techniques usually depend on acquisition specifics such as image quality, resolution, head pose and illumination conditions.
In the contour-based lip motion representation, only motion vectors computed on the pixels along the extracted lip contour are taken into account and the rest is discarded. In this case, the two sequences of x and y motion components on the contour pixels are separately transformed using one-dimensional DCT. Note that the length of the resulting sequence of motion components on each direction may vary from one frame to another according to varying lip shape. In order to obtain a feature vector of fixed size in each frame, prior to 1-D DCT transformation, the length of the sequence is normalized to a fixed number by using linear interpolation. This number is the maximum number of contour points achieved in any lip frame of all available sequences. The DCT coefficients are computed separately for and directions are concatenated to form the feature vector [10].

2.4 FUSION AGENTS

In this paper we have proposed a framework in which we have used agents. An agent is an autonomous entity, which gets information from the environment and processes it to reach a decision about further actions [4]. Agents are small software entities, which can be equipped with reasoning, learning and communication skills and display goal-oriented behavior [5].
In my proposed framework, first of all the speaker is identified as a correct person with the help of face recognition and speech recognition. This may be used as an interface for legal identification of a user.
In this we have two agents for multimodal inputs. First being audition processing agent which deals with audio data. This agent collects and comprehends the hearing information. Second, is visual processing agent which takes into account two types of data. One related with lip and the other related with face detection. So, this visual processing agent deals with two different types of multimodal inputs. Lip tracking and detection and face detection is done and the information is passed over to visual processing agent. This agent basically collects data from these inputs and comprehends the visual information. This bimodal data proves very useful in giving visual information.
After the collection of data by the above two agents, the Information fusion agent finally comprehends each of these two agents and fuses data given by these two agents so that required information is properly passed and speech is recognized. Since, we have combined visual information with audio information, the automatic speech recognition becomes fast and is better than if we would have considered either of the information alone.
On the basis of proposed framework we conducted pilot experiments that contributed to the design of frame representation for our task domain and yielded preliminary test data to evaluate the system.

III. CONCLUSION

This paper addresses a unique framework in which different agents have been introduced. These software agents help in human computer interaction. By using audio, video and lip motion, information can be conveyed to or exchanged with computer. The audio and visual agents process the audio and visual information, which finally is fused by audio visual fusion agent. The audio agent and visual agent act as intermediary agents. This audio visual fusion agent finally helps in automatic speech recognition. The agent concept is a new one and these agents assist humans. They also decrease the workload of the humans and help in a better interaction with computers. They also provide a more natural environment for interaction of the human and computer.

REFERENCES

[1] Ergin Erzin, Yucel Yemez, A.Murat Tekalp, "Multimodal speaker identification using an adaptive classifier cascade based on modality reliability" IEEE Transactions on Multimedia, vol. 7, no. 5, pp. 840-852, October 2005.

[2] Maja Pantic, Leon J. M. Rothkrantz, "Toward an Affect-Sensitive Multimodal Human–Computer Interaction**"**, Proceedings of the IEEE, vol. 91, no. 9, pp. 1370-1390, September 2003.

[3] Engin Erzin, Y¨ucel Yemez, A. Murat Tekalp, "Multimodal Speaker Identification using Adaptive Decision Fusion with Reliability Weighted Summation", COST278 and ISCA Tutorial and Research Workshop, on Robustness Issues in Conversational Interaction, University of East Anglia, Norwich, UK, August 30-31, 2004.

[4] Maximilian Kruger, Achim Schafer, Andreas Tewes, Rolf P. Wurtz, "Communicating Agents Architecture with Applications in Multimodal Human Computer Interaction", Informatik 2004.

[5] Elfriede I. Krauth, Jos van Hillegersberg, Steef L. van de Velde, "Agent-based Human-computer-interaction for Real-time Monitoring Systems in the Trucking Industry**"** IEEE Proceedings of the 40th Hawaii International Conference on System Sciences pp. 1-7, 2007.

[6] Zhen Zhu, Jing-Yan Wang, "Multi-Agent Based Approach To Support HCI", IEEE, Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, pp.188-191, 13-16 August 2006.

[7] H.E. Cetingul, E. Erzin, Y. Yemez, A.M. Tekalp, "Multimodal speaker/speech recognition using lip motion, lip texture and audio", Signal Processing vol. 86 pp. 3549–3558, December 2006.

[8] Fengjun Chen, Zhiliang Wang, Zhengguang Xu, Yujie Wang, Fang Liu, "A Facial Expression Recognition Algorithm based on Feature Fusion", IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application, pp. 381-385, 2008.

[9] Erno Makien, "Face Analysis Techniques for Human-Computer Interaction", Dissertations in Interactive Technology, No. 8, Tampere 2007.

[10] H. Ertan Çetingül, Yücel Yemez, Engin Erzin, and A. Murat Tekalp, "Discriminative Analysis of Lip Motion Features for Speaker Identification and Speech-Reading" IEEE Transactions on Image Processing, vol. 15, no. 10, pp. 2879-2891, October 2006.

[11] L. Deng, Y. Wang, K. Wang, A.Acero, H. Hon, J. Droppo, C. Boulis, , M. Mahajan, and X.D.Huang, "Speech and language processing for Multimodal Human Computer Interaction", Journal of VLSI Signal Processing vol. 36, no. 2-3, pp. 161–187, 2004.

[12] Zhihong Zeng, Jilin Tu, Brian M. Pianfetti, Jr., and Thomas S. Huang, "Audio–Visual Affective Expression Recognition Through Multistream Fused HMM", IEEE Transactions On Multimedia, vol. 10, no. 4, pp. 570-577, June 2008.

AUTHORS PROFILE

Prof. S.Q. Abbas is Ph.D and currently he is working as a Director , Ambalika Institute of Management and Technology, Lucknow, India. He has obtained his MS(CS) from BITS Pilani. His research area is Software Engineering, Data Mining and Human Computer Interaction. Dr. Abbas has published many of the valuable research papers in various national and international journals.

Nidhi Srivastava is M.Phil(CS), MCA and pursuing Ph.D (CS) from UP Rajarshi Tandon Open University, Allahabad. She is currently working as Sr. Lecturer in Amity University. Her research area is Human Computer Interaction (HCI).