IOSR Journal of Computer Engineering (IOSRJCE)
ISSN: 2278-0661 Volume 5, Issue 1 (Sept.- Oct. 2012)PP 31-36
www.iosrjournals.org

Advances in Automatic Speech Recognition: From Audio-Only To Audio-Visual Speech Recognition

Akriti Bahal[1], Rachit Gupta[2]

[1],[2]*(Computer Science Department, Maharaja Agrasen Institute of Technology, Guru Gobind Singh Indraprastha University,Delhi-110086,India)*

Abstract: *Major developments in the field of finding more natural ways of interacting with computers have been taking place. The clear focus lies on making technology more approachable to people. The concept that computers can comprehend our various gestures by eyes, voices, touch and our different movements to interact is called the Natural User Interface (NUI). Today, many of these elements are available in mobile phones, PCs, and in other devices. Speech technologies particularly play a substantial role in this evolution. Significant advancement in automatic speech recognition (ASR) for well defined applications like dictation and medium vocabulary transaction processing assignments in comparatively controlled environments have been made. But, automatic speech recognition still has to reach a level needed for speech to become a completely pervasive user interface because even in clean acoustic surroundings, the state of ASR system performance falls behind human speech perception. Visual speech recognition, however, is a promising source of extra speech information and it has successfully exhibited to enhance noise robustness of automatic speech recognizers, thereby promising to expand their usability in the human computer interaction. In this paper, the main components of audio-visual speech recognition, i.e., the audio and the video components are discussed, along with the latest advancements made in this field. Further, the research goes beyond the recent advancements and discusses the future scope of audio video speech recognition and mentions some likely future developments, evaluating each on the basis of its performance. Graphs are plotted based on experiments to depict the performance improvements from audio-only ASR to audio-video ASR, along with its expected performance level in future.*

Keywords - *Audio-Video Speech Recognition, Automatic Speech Recognition Advancements, Lipreading, Speech Recognition.*

I. Introduction

The first advancements in speech recognition antedate the invention of the modern computer by greater than 50 years. Alexander Graham Bell was interested in an experiment to transmit speech by his wife, who was deaf. He earlier thought to make a device which would convert audible words into a visible picture that a deaf person could construe. He was successful in producing spectrographic images of sounds, but his wife was unable to interpret them. This line of study ultimately led to his invention of the telephone. For many decades, scientists created experimental procedures of computerized speech recognition, but the computing power available at that time restricted them. It was only in the 1990s that computers powerful enough to handle speech recognition became available to the normal consumer. We have made a large amount of progress in automatic speech recognition (ASR) since then. It is widely used in well-defined applications present in most of the mobile phones and operating systems. Applications like dictation and medium vocabulary transaction processing tasks are largely used today. However, ASR performance has yet to attain the level needed for speech to become a completely pervasive user interface because even today, the state of the ASR system performance lags human speech perception by up to an order of magnitude even if clean acoustic environments are considered. Also, the current systems are quite sensitive to speech variations. But, non-traditional approaches, that make use of orthogonal sources of data to the acoustic input, are being developed to achieve ASR performance near to the human speech perception level, and robust enough to be used in field applications. Visual speech is the most promising source of supplementary speech information, and it is also not affectedbytheacousticenvironmentandnoisedisturbances. Human speech perception is bimodal in nature: Humans combine the available audio and visual information in determining what has been spoken, specifically in noisy environments. The visual modality greatly increases the speech intelligibility in noise. In addition, bimodal fusion of audio and visual stimuli in deciphering speech has been shown by the McGurk effect. For instance, when the spoken sound /ga/ is superimposed on the video of a person speaking /ba/, most of the people decipher the orator as speaking the sound /da/. Furthermore, visual speech is of specific importance to the people who are hearing impaired. Mouth movement plays an essential role in both sign language and concurrent communication between the deaf. The hearing impaired people lip read well, and probably better than the generalpopulace. There are three prominent reasons why vision is advantageous to human speech perception: It aids speaker localization, it comprises of speech segmental data that complements the audio, and it also provides other supplementary information about the place of utterance. The latter part is due to partial or absolute visibility of speech organs, i.e., tongue, teeth, and lips. Place of articulation information can help remove the confusion from confusing sounds that arises when only acoustics are considered. Jaw and lower face muscle movement is correlated to the produced acoustics and its visibility enhances human speech perception. The reasons have stimulated substantial interest in automatic recognition of visual speech, originally known as automatic lipreading, or speechreading. Work in this field is targeted towards enhancing ASR by making use of the visual modality of the talker's mouth region along with the traditional audio modality, leading to audio-visualautomaticspeechrecognitionsystems. Incorporating the visual modality has depicted to outperform audio-only ASR over a wide range of conditions. These performance gains are specifically substantial in environments with noise, where conventional acoustic-only ASR performs badly. Along with the reducing cost of quality video capturing systems, these cases make automatic speechreading a greatly robust ASR system. In this paper, the performance benefits brought in by incorporating the visual component in ASR are discussed, along with the advancements in the applications that are developed using the same. In addition to this, the expected future developments of ASR are discussed, and graphs are plotted based on the experiments to show the performance improvements from audio-only ASR to audio-visual ASR (AV-ASR), along with its expected level in future.

II. Relatedwork

Automatic speech recognition (ASR) is seen as an important part of human-computer interfaces thatareenvisaged to use speech, among other means, to attain natural, pervasive, and omnipresent computing. But,although ASR has experienced weighty development in specifically defined applications like commands through speech, dictation and mediocre vocabulary transaction processing jobs, its performance has yet to achieve the level needed for speech to become a completely widespread user interface. The state of ASR lacks robustness to channel and environment noise continues to be a major impediment. Also, the system performance lags human speech perception. Therefore, this clears the air for the need of non-conventional approaches, which use the information orthogonal to the audio input to achieve ASR performance closer to the human speech perception level that is strong enough to be used in field applications. Due to this reason, visual speech forms a promising such source, which is not hindered by acoustic noise.

2.1AudioComponent

The audio component involves the basic speech recognition process. Speech recognition is the translation of conversational words into textual information. It is known as automatic speech recognition or computer speech recognition as well. In some speech recognition systems, the process of training is used where the speaker reads portions of text into the system. These systems construe the speaker's voice and employ it to fine tune the recognition of that person's oration, which thereby results in more precise transcription. While there are other systems that do not use training, and they are called speaker independent systems.

Audio speech recognition applications comprise voice user interfaces such as speed dialling (for example "Call Office"), call routing (for example, "I want to make a conference call"), domotic appliance control, search (for example, "search for a nearby restaurant"), simple data entry (for example, entering a phone number), and speech-to-textprocessing(forexample, wordprocessors or emails).

The performance of speech recognition systems is generally assessed in terms of accuracy and speed. Accuracy is generally evaluated with word error rate (WER), while speed is calculated with the real time factor. Among the other measures of accuracy are Single Word Error Rate (SWER) and Command Success Rate (CSR). Although, audio speech recognition has a wide range of applications, but it is a very involved problem. Vocalizations differ in terms of accent, pronunciation, clarity, roughness, tone, pitch, volume, and speed. In addition to this, speech is also warped by background noises and disturbances. Accuracy of speech recognition also varies with other factors such as vocabulary size and confusability, speaker dependence versus independence, task and language constraints and read versus spontaneous speech.

In order to eliminate these problems in accuracy the video component was introduced in automatic speech recognition, which involves image processing abilities in lip reading to help speech recognition systems in recognizing nondeterministic phones or giving superiority amongst near probability decisions.

2.1.1 Applications

- iPhone'sSirifeature Apple's introduction of Siri for the iPhone 4S greatly increased the role of speech-activated assistant technology. Siri is an smart personal assistant and knowledge navigator that functions as an application for Apple's iOS. The application employs a natural language user interface to respond to questions, provide suggestions, and carry out actions by commissioning requests to a set of Web services. In addition to this, Siri provides a natural, conversation-like experience with speech-activated assistants. For example, to use Siri there is no need to memorize a strict list of commands, a person can simply speak as if he were to speak to another person without having to adhere to a special speaking syntax. But there are also certain limitations to the Siri technology. Siri often misconstrued casually spoken orders, making it easier in many situations to carry out the tasks manually.

- GoogleChrome'sVoiceSearchfeature Voice Search is a Google Chrome extension that allows the user to perform search using his voice. It is not developed by Google, but it employs an experimental Chrome feature known as form speech input. The feature is employed by default in the dev channel builds, but it can also be manually employed by the addition of a command-line flag. Voice Search includes a number of default services like Google, Wikipedia, YouTube, Bing, Yahoo, DuckDuckGo and Wolfram|Alpha. In addition to this, we can also add our own user-defined search engines. Also, it includes a speech input button for all websites that use HTML5 search boxes. This extension necessitates a microphone. In spite these positive aspects; it includes certain demerits as well. One of them include that speech input is very experimental and may infer wrong data if not spoken clearly.
- Speechenhancedcommenting Inefficient and inadequate commenting of code is a principal problem in software development. Coders find that commenting consumes their substantial amount of time and is also frustrating. The general scenario followed by programmers is that under most circumstances they write a section of code and return later to comment, this substantially reduces the accuracy of the information. Speech Enhanced Commenting is a scheme that takes vantage of an unused human output - voice. With the help of this system, the user is allowed to speak their comments as they program. This process is natural and can take place simultaneously as the user types the code. Therefore, no extra time is required towritecomments.There are even more advantages beyond providing a stage that promote more thorough commenting. By having audio clips for every line, functions can be read out loud, line-by-line, and this can prove to be very helpful if a user in unacquainted with the code and desires to review. Although, the fact that the audio clips can be read like a story, succeeding the execution path of the program, can be substantially helpful for debugging. Also, as the users can concentrate on the code and listen to the description, this provides a higher information throughput than the one provided by just reading alone.

2.2VisualComponent

The sight of a speaker's face has been known for long to improve speech comprehensibility, mainly where the acoustic speech signal is deteriorated by noise and disturbances or in cases where there is hearing-damage. The advantage obtained from the visual, facial signals has been quantitatively measured to be tantamount to a rise of 8-10 dB in the signal-to-noise ratio when speech sentences are presented in a noisy background. This rise in the signal-to-noise ratio depicts that if the acoustic inputs to traditional speech recognition systems would be increased by data about the visible speech gestures, it would result into an improved-performance, audio-visual recognition system. The incorporation of visual data into speech recognition offers a substantial degree of acoustic noise immunity, which the traditional speech recognition systems could not offer, at a moderate computationalcost. With the inclusion of a visual component in speech recognition, comes the management of image data, which can easily be achieved by recognizing and abstracting features from the images and using them instead. In order to identify only the germane features and to reduce redundancy in the information, an approach to use the images themselves as the source of data and to adopt a statistically oriented, data driven approach is used. The benefit of using the images is that they enclose not only the lips, but other perceptually importance features such as the tongue, teeth, mouth, and skin texture as well.

2.2.1Faceandfacialfeaturesextraction

Fce and facial part detection has drawn substantial interest, but at the same time it constitutes a difficult problem, most significantly in cases where the background, head pose, and lighting are changing. Many such systems use conventional image processing methods, such as color segmentation, edge detection, image thresholding, template matching, or motion data, whereas in other systems a statistical modeling approach is used, which employs a neural networks for example. For the systems belonging to the second category, for a video frame, face detection is first carried out by searching for face candidates that contain a comparatively high percentage of skin-tone pixels. Once a face has been detected, a mixture of facial feature detectors is employed to calculate the positions of 26 facial features, which include the lip corners and centers. Even though mouth width and the amount that the mouth opens while speaking are known to be perceptually important, the catalog of apposite features is in general difficult to specify completely and some of the important information such as the tongue, which may look only as an indefinite and ill-defined object, cannot easily be described in simple parametric terms. Nevertheless, still a number of visual speech recognition systems continue to use a range of characteristics of this type with some amount of success.

2.2.2Usingdirectimages

Early experiments on vowel identification which had limited perceptual range suggested that important visual hints could substantially be held in monochrome, dynamic recordings of the oral area of the individual talker's face in which the spatial resolution was as low as 16 x 16 pixels. This indicated two things. First, that the visual cues to vowel identity could be held in images of comparatively low spatial resolution and second, that the visual hints to vowel identity could be captured in a low-dimensionality intermediate presentation of the images. Using direct images has important implications for data compression of spoken images because it suggests the likelihood of an efficient image encoding scheme.

2.2.3LipContourTracking

Lip Contour Tracking involves tracking the mouth region, and there are a number of algorithms to do the same. Some famous procedures include snake, templates, and active shape and appearance models. A snake is an elastic curve presented by a given number of control points. The control point coordinates are repeatedly updated by meeting towards the local minimum of an energy function, determined on the principle of curve smoothness constraints and a matching measure to desired features of the image. Another popularly employed method for lip tracking is by means of lip templates. Templates comprise of parameterized curves which are adapted to the desired shape by minimizing an energy function, determined in a similar fashion to snakes. Active shape and appearance models develop a lip shape or ROI appearance statistical model. This presumes that, given small disturbances from the actual adjustment of the model to the image, a linear relationship is present between the difference in the model projection and image, and the needed updates to the model parameters. An iterative algorithm is employed to adjust the model to the image data. These models form the basis of lip reading, which are further employed for various applications.

2.3Audio-VisualIntegration

Audio-visual integration intends at merging the two available speech informative streams into a bimodal classifier with higher performance to both audio- and visual-only recognition. Different information fusion algorithms have been considered for audio visual automatic speech recognition, which differ in their fundamental design, the speech classification technology used, and in the terminology adopted. The conventional hidden Markov model (HMM) based approach for ASR uses acoustic-based speech classes and Gaussian mixture densities as the class-conditional probabilities of the feature observations of interest. Other than this, there are a number of feasible alternatives such as hybrid HMM/neural network, or support vector machine based ASR architectures, perhaps using visual based speech classes as well.
The audio-visual integration techniques are grouped into two categories feature fusion and decision fusion methods. The first is founded on training a single classifier, i.e., of the similar form as the audio and the video only ones, on the linked vector of audio and visual features, or on any relevant transformation of it. On the other hand, decision fusion algorithms employ the two single-modality (audio- and visual-only) classifier outputs to identify audio visual speech. Generally, this is attained by linearly combining the class-conditional observation log-likelihoods of the two classifiers into a common audiovisual score, using relevant weights that capture the dependability of each single-modality data stream. Besides these categories, there exist methods that incorporate characteristics of both. One such hybrid fusion method is the one in which the representation of all techniques at first presumes an "early" temporal level of audio-visual integration, namely at the HMM state. Such models are called asynchronous models.

2.3.1 Applications

- **MAVIS**

  The Microsoft Research Audio Video Indexing System (MAVIS) is a system of software elements that employ speech recognition technology to allow searching of digitized spoken content, be it from meetings, conference calls, voice mails, presentations, online lectures, or even Internet video. Another advantage of MAVIS is the ability to produce automatic closed captions and keywords that can enhance accessibility and discoverability of audio andvideofileswithspeechcontent. Microsoft's previous products on speech recognition technology include Windows 7, TellMe.com, Exchange 2010, and Office OneNote. MAVIS adds to the catalogue of Microsoft applications and services that employ speech recognition. MAVIS is designed to permit searching of 1000s or even 100,000s of hours of conversational speech with various talkers on various topics. With this, the user can type a search term or phrase and receive the links to where those words were conversed.
- **RealTimeLipTracking**

Advancements in dynamic outline tracking allow scarce representation of the outlines of moving contours. Because of the rising computing power of general-purpose workstation, it is now possible to track human faces and portions of faces in real-time without additional hardware. Two such lip trackers are the ones that tracks lips from a profile view and the other from a front view. These were employed to abstract visual speech recognition features from the lip contour. In both the cases, visual features are subsumed into an acoustic automatic speech recognizer. Tests on moderate isolated-word vocabularies that employ a dynamic time warping based audio-visual recognizer manifest that real-time, outline-based lip tracking can be used to supplement acoustic-only speech recognizers permitting strong recognition of speech in the presence of acoustic noise. For instance, ``m" and ``n" which are letter capable of being confused acoustically, especially in noise situations and in situations where disturbances are high, are easy to differentiate visually, that in ``m" lips close at beginning, where as in ``n" they do not.

III. Performancecomparison

Experiment was performed to track the performance of audio-only, visual-only, and audio-visual integrated automatic speech recognition systems when different levels of simulated noise were added to the acoustic signal. Percentage word error rate was calculated for each value of the simulated noise. When an acoustic signal to noise ratio (SNR) of 23 dB was considered, the audio-only system showed a word error rate of 15.3%, while the audio-visual integrated system showed 1.6% of word error. Similarly, for an SNR of -13 dB, audio-only recognizer showed a word error as high as 100%, while that of audio-visual recognizer was only 22.3%. The visual-only recognizer maintained a constant word error rate of 23% for all noise levels. The performance of audio-visual integrated recognizers was found to be better than that obtained from either of the acoustic-only or the visual-only recognisers at virtually all acoustic noise levels. The graph below depicts the result of the experiment showing the percentage word error rate of each of the three recognizers at different acoustic noise levels.

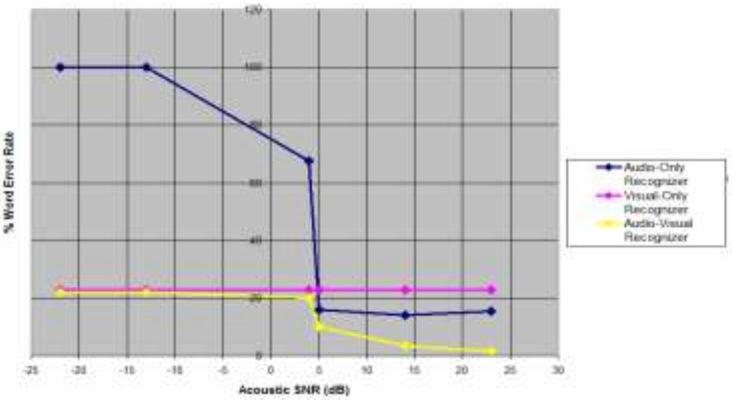

Fig 1: Performance Comparison between Audio-Only, Visual-Only and Audio-Visual Recognizers

Furthermore, the expected improvements in performance for future audio-visual speech recognizers were theoretically estimated based on the expected advancements in these recognizers. In future, speech recognition may become speech understanding. The statistical models that permit computers to decide what a person said may someday permit them to even comprehend the meaning behind the spoken words. Though it is a great leap in terms of computational power and software sophistication, but it is likely to take place in future. With the development of this, the disturbances caused due to noise will reduce even further. On the basis of these expected advancements, the graph below shows the performance comparison between the present day audio-visual speech recognizers and future speech recognizers.

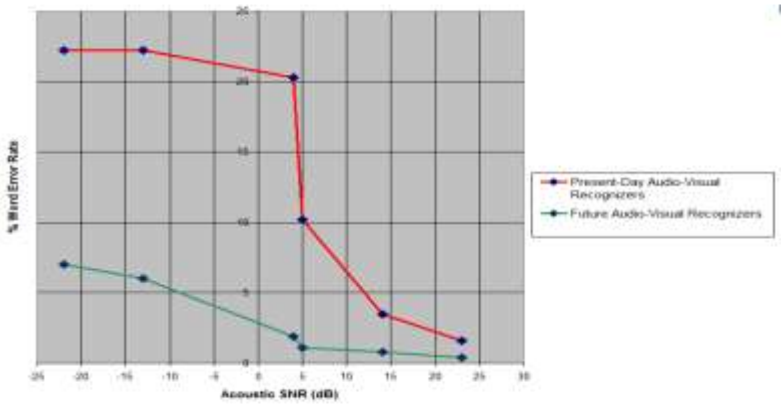

Fig 2: Performance Comparison between Present-Day AV Recognizers and Future AV Recognizers

IV. Futurework

Large amount of developments have been made in the field of audio-visual speech recognition, and it still hold a large scope for further advancement. Although speech recognition is already included into products like Microsoft's OfficeXP, large numbers of users still prefer to use their keyboards. Speech recognition can be trained to recognize specific speaker's voice. However, construing sounds from different speakers can be evenmore challenging, unless a library of sounds is employed. Still, even though speech recognition by a computermaybefarfromperfect,thefutureisbright.The IBM's superhuman speech recognition, aims to get a performance analogous to humans in the next five years.

Translating a command, for example "Play 92.3" requires the device to grasp the fundamental context of the command, a characteristic known in Embedded ViaVoice 4.4 as Free-Form Command. For Free-Form Command to work successfully, the scheme should recognize two things, namely, first that the user is referring to the radio, even if he does not make use of the actual term "radio", and secondly, the software needs to be programmed to comprehend that the term "play" is also a command to tune the radio to the desired station. In the future, speech recognition is expected to become speech understanding. Today, we can talk to computers, but in 25 years, they may talk back as well.

V. Conclusion

The implementation of experimental systems for automatic speech recognition has already shown the advantages of adding a visual component to the traditional, acoustic systems. Cognitive models can provide a beneficial conceptual framework in the quest for recognition architectures in which the audio and visual components are optimally integrated, as the present work is depicting.

Furthermore, including the visual component, not only enhances the performance of the automatic speech recognizer under disturbances by noise, but it also expands the usability of automatic speech recognizers in the field of human computer interaction.

In this paper, experiments were performed to find out the performance variations between the traditional audio-only speech recognizer and audio-visual speech recognizer. The results depicted that at almost all noise levels, audio-visual speech recognizers greatly outshone the conventional acoustic recognizers. Further, theoretical analysis of the future of audio-visual speech recognizers was also performed and the estimated variations with signal to noise ratio were noted.

The future of automatic speech recognizers is very bright, and a large leap of advancement in audio-visual recognition is likely in future.

VI. Acknowledgement

I would like to thank Mrs. Ruchi Goel, Assistant Professor, CSE Department, Maharaja Agrasen Institute of Technology, IP University, Delhi, India, for her guidance and support during the research work.

References

[1]. M . Jackson, *Automatic Speech Recognition: Human Computer Interface for Kinyarwanda Language*, MakerereUniversity,2005.

[2]. B.Chen, *Recent Developments in Speech Recognition Technologies and Their Applications*, NationalTaiwanNormalUniversity,2003

[3]. N. Michael Brooke, *Using the Visual Component Automatic Speech Recognition*, University of Bath, Bath,UK.

[4]. G. Potamianos, C. Neti, G. Gravier, A. Garg, A.W. Senior, Recent Advances in the Automatic Recognition of Audiovisual Speech, *Proc. of the IEEE, Vol. 91, No. 9, 2003.*

[5]. G. Potamianos, C. Neti, J. Luettin, I. Matthews, Audio-Visual Speech Recognition: An Overview, *Issues in Visual and Audio-Visual Speech Processing* (MIT Press, 2004).

[6]. J. Huang, Z. Liu, Y. Wang, Y. Chen, and E. Wong, Integration of multimodal features for video scene classification based on HMM, *Proc. Workshop Multimedia Signal Processing, 1999, pp. 53–58.*

[7]. G. Gravier, S. Axelrod, G. Potamianos, and C. Neti, Maximum entropy and MCE based HMM stream weight estimation for audio-visual ASR, *in Proc. Int. Conf.*