

A Vision-based Microphone Switch for Speech Intent Detection

Giridharan Iyengar and Chalapathy Neti

Human Language Technologies
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598.
{giyengar,cneti}@us.ibm.com

Abstract

In this paper we present our system for speech intent detection. In traditional desktop speech applications, the user has to explicitly indicate intent-to-speak to the computer by turning the microphone on. This is to alleviate problems associated with an open microphone in an automatic speech recognition system. In this paper, we use cues derived from user pose, proximity and visual speech activity to detect speech intent and enable automatic control of the microphone. We achieve real-time performance using pre-attentive cues to eliminate redundant computation.

1 Introduction

Speech recognition systems have been a major focus of ongoing research into an intuitive and natural Human-Computer interaction. Significant progress has been achieved in terms of accuracy of automatic speech recognition (ASR) to the point that commercial systems are being deployed regularly, especially in desktop and telephony applications.

In a desktop application, while the ASR performance is reasonable, the user of such a system has to explicitly turn the microphone ON or OFF indicating intent to speak. This is needed to prevent the ASR system from trying to recognize background noise. Contrast this with a telephony channel where the intent to interact is established as long as the connection is kept alive. In addition, telephone based ASR systems perform silence detection to mitigate the open mic prob-

lem during an interaction. While such a scheme can be attempted for desktop systems, the number of instances where speech sounds are produced but not intended for the ASR system are many more. For example, the user might be talking with someone else in the room or be interacting with an ASR system on the telephone or even with another computer.

In Human-Human communications, one prominent modality that is used in addition to speech is the visual modality. For example, we attract a person's attention by making eye contact, being physically proximate before engaging in a conversation. In addition, we can see the lips move in addition to hearing the speech. In short, humans use visual cues to establish intent to speak. Researchers have experimented with such cues in Human-Computer dialog turntaking with the computer directing its gaze away from the user and to indicate to the user that the computer is busy[1]. Apart from establishing intent to speak, humans use visual speech cues to better understand speech[2]. And there have been some success in integrating visual cues in an ASR system [3, 4, 5]. Joint processing of audio and visual information have been used successfully in speaker change, speaker identification etc as well[6, 7].

In this paper, we propose to use the visual channel for establishment of speech-intent. One can argue that rather than using the visual channel, a user can explicitly address the device with a unique name to establish intent. This has a couple of obvious problems. As the number of devices increase in our environments, remembering their names becomes a cumbersome task. In addition, requiring a user to refer to an object in or-

der to establish intent takes away from the naturalness of the interaction. In addition if the user is simultaneously interacting with multiple devices, something that is fairly common even now, it becomes difficult to disambiguate the device to which a particular utterance was directed without explicit naming of the device in each utterance. This situation is clearly undesirable.

In this paper, we concentrate on establishment of intent. We assume that the Human-Computer interaction is a traditional desktop *speech-in, visual-out* scenario. Specifically, we define intent to speak in the above scenario as a face close and frontal with respect to the computer screen, with lips moving. The intent is established as soon as a frontal face is detected and is maintained as long as the lips are moving in addition to a frontal face being detected.

2 Frontal face and feature detection

We employ the face detection system presented in [8, 9] as our initial face detector. Briefly, a skin-tone segmentation is performed to locate image regions where colors indicate presence of a face. These regions are then scored with a combination of a Fisher Discriminant and a Distance From Face Space (DFFS) to give a face likelihood score. Higher the total score, the higher the chance that the considered region is a face.

A similar method is applied, combined with statistical considerations of position, to detect the features within a face. Notice that this face and feature detection scheme was designed to detect strictly frontal faces only, and the templates are intended only to distinguish strictly frontal faces from non-faces: general facial poses are not considered at all.

The face score is then compared to a threshold to decide whether or not a face candidate is a real face. This score, for a given user, varies almost linearly as the user is turning her head — the score being the highest when the face is strictly frontal and the lowest when it is in profile. But this face score is user-dependent and we could not find a user independent

threshold that could have allowed us to decide on the frontalness of the pose by simple thresholding of the face score.

2.1 Geometric pruning for user-independent pose detection

In order to get a user-independent pose estimate we perform additional pruning of the face candidates using the feature detectors. We allow a lower threshold for face detection which increases the number of non-frontal faces being detected along with higher false alarms. Once a face candidate is identified and facial features of interest are located, we prune candidates and estimate the frontality of pose based on the following aspects:

- The sum of all the facial feature scores is compared to a threshold to decide whether the candidate should be discarded or not.
- The number of main features that are well recognized is used to further prune out unreliable matches. We discard candidates with low score for the eyes, the nose and the mouth.
- The ratio of the distance between each eye and the center of the nose is estimated. If the pose is frontal, this ratio will be unity.
- The ratio of the distance between each eye and the side of the face region is also expected to be unity for frontal poses (each face candidate is normalized so that the side of the detected face square corresponds to the eye separation [8]).

The last two ratios will differ from unity for non-frontal faces. So, we compute these ratios for each face candidate and compare them to 1 to decide whether the candidate has to be discarded or not. Then, if one or more face candidates remain in the candidates stack, we will consider that a frontal face has been detected in the considered frame. In addition to this, we allow a burst parameter that allows a decision to remain static for a certain number of frames. This is to prevent the microphone from being accidentally turned off because of occasional false negatives. Further details of this algorithm can be found in

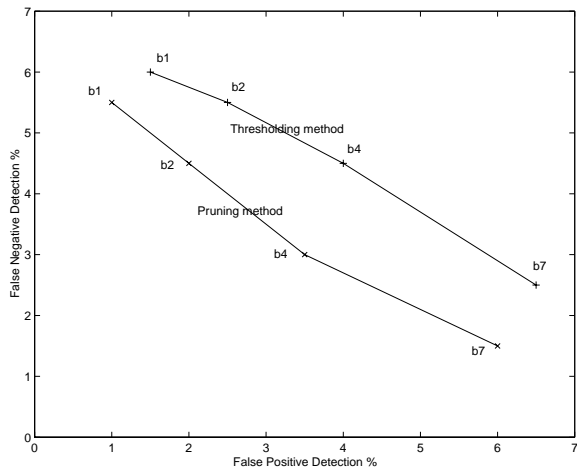


Figure 1: ROC curve for the pruning method compared with simple face score thresholding. The different datapoints correspond to different burst values.

Ref.[10]. Figure 1 shows the comparison between the pruning method and simple thresholding for a small dataset of 10 users.

3 Pre-attentive cues for real-time performance

It may not be necessary to detect face pose in every incoming frame of video. For example, if there is no face in front of the camera, there is relatively little change in the incoming image. Similarly once a person is frontal and speaking in front of the camera, the head motion is small and once again there is little change in the incoming image. Only the mouth shape changes with speech. We try to exploit this by performing a simple image difference operation right at the outset and invoking the face detector/tracker only for large changes in the image. This is akin to the reptilian vision which detects objects only if they move. The main advantage of using such pre-attentive cues is computational savings at very little sacrifice in per-

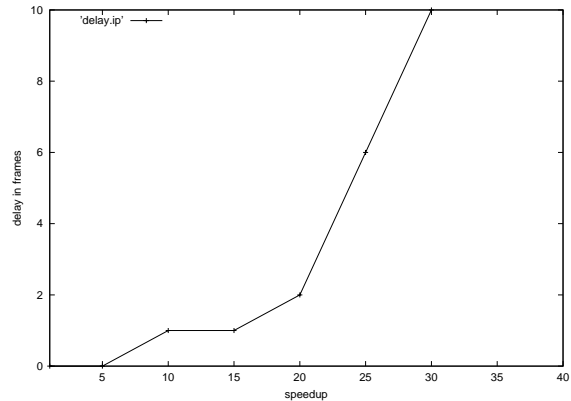


Figure 2: Pose detection delay vs. speedup

formance. Since we have not changed the underlying detection or pruning algorithms, the detection performance remains the same for static test sets. In dynamic test sets (where the user pose is slowly changing from frontal to non-frontal and vice versa), we expect a small delay in detection of the onset of a pose and likewise a delay in detection of the end of a frontal pose. In addition, this delay varies with the pre-attentive threshold. For large values of this threshold, no change in the incoming image is big enough and therefore no pose will be detected. For small values of this threshold, we gain computational savings at the cost of a small delay in the detection. An experiment with 4 subjects changing their pose from non-frontal to frontal to non-frontal continuously is shown below in Fig. 2. The Y-axis shows the delay in number of frames between the onset of a frontal pose in our hand annotated ground-truth and the pose being detected by the algorithm. The X-axis shows the corresponding processing speedup achieved. We notice that for small thresholds, the delay is quite small but the computational speedups are large. We select 15x speedup as our operating point in the final system. This is sufficient to give us roughly 30 frames per second performance with a reasonable desktop computer.

4 Detection of Visual Speech Activity

The final component of our system is the detection of visual speech activity. This is used to validate a speech intent decision triggered by a frontal pose. Once a frontal pose is detected, the microphone is switched ON and the ASR system is activated. However, the user may not be speaking and we encounter the open mic problem as before. To counter this, we employ a similar technique to pre-attentive cues to determine if the user has speech intent. From the facial feature estimates, we extract a bounding box encompassing the lips. This bounding box is compared with the corresponding box in the previous frame. We note here that this bounding box is extracted irrespective of whether we perform face detection in a particular frame. For the frames where we do not perform face detection, the feature locations estimates from the last detection are used to extract the bounding box. We compare the average illuminance of the extracted mouth regions. The intuition is that when the mouth is open, the average illuminance of the bounding box is lower and likewise it is higher when the mouth is closed. Therefore to detect speech, it is enough to detect *change* in average illuminance of this bounding box and threshold it. Ofcourse, it is easy to fool the system by dropping one's jaw and simulating speech activity. This can be remedied by combining this method with an audio based speech/silence detector. We have done some initial work in visual speech/silence detection for classifying each frame as a speech, silence or a non-speech frame[10]. However, for controlling the microphone and the ASR system as in this paper, it is not enough to classify frames in isolation but rather determine if speech is occurring (as opposed to say, yawning). This is a hard problem and as of now we do not have a real-time implementation. We therefore implement the current simplistic technique of image differencing to detect visual speech activity.

It is known from previous perceptual studies that the onset of visual speech activity precedes the audio onset of speech. Bregler et al, in their work on audio-visual speechreading, found that the visual channel precedes the audio channel by approximately 120ms[11]. While in theory it is possible to exploit

this onset delay in our application, we find that other issues such as operating system delays make it difficult to get a reliable estimate of the onset delay and exploit it.

We have begun work on combining speech intent detection with audio-visual speaker identification. This is to enable authentication in addition to enabling a natural interface. While the work is promising, it is at an early stage to report on the performance.

5 Summary

In this paper we presented a vision based microphone switch for desktop speech applications that uses face detection, pre-attentive cues and visual speech activity. This enables a more natural speech interface. Such a system can be combined with audio and visual speaker identification to perform authentication in addition to enabling a natural interface.

References

- [1] J. Cassell, T. Bickmore, L. Campbell H. Vilhjmsson, and H. Yan, "Human conversation as a system framework: Designing embodied conversational agents," in *Embodied Conversational Agents*, Cassell, J. et al., Ed., Cambridge, MA, 2000, MIT Press.
- [2] Q. Summerfield, "Some preliminaries to a comprehensive account of audio-visual speech perception," in *Hearing by Eye: The Psychology of Lip-reading*, B. Dodd and R. Campbell, Eds., London, 1987, pp. 3-51, Lawrence Erlbaum Associates.
- [3] G. Potamianos, A. Verma, C. Neti, G. Iyengar, and S. Basu, "A cascade image transform for speaker independent automatic speechreading," in *Intl. Conf. Multimedia and Expo*, New York, 2000, vol. II, pp. 1097-1100, IEEE.
- [4] C. Neti, G. Potamianos, J. Leuttin, I. Matthews, H. Glotin, D. Vergyri, J. Sisson, A. Mashari,

and J. Zhou, "Audio-visual speech recognition," CLSP Summer Workshop Tech. Rep. WS00AVSR, Johns-Hopkins University, Baltimore, MD, 2000.

- [5] Tshuan Chen and Ram R. Rao, "Audio-visual integration in multimodal communication," *Proc. IEEE*, vol. 86(5), pp. 837–852, May 1998.
- [6] G. Iyengar and C. Neti, "Speaker change detection using joint audio-visual statistics," in *RIAO*, Paris, France, April 2000.
- [7] Benoit Maison, Chalapathy Neti, and Andrew Senior, "Audio-visual speaker recognition for video broadcast news: some fusion techniques," in *IEEE Multimedia Signal Processing (MMSP99)*, Denmark, September 1999.
- [8] Andrew W. Senior, "Face and feature finding for a face recognition system," in *Intl. Conf. Audio-Video based Bio. Pers. Authen.* 1999, Lecture Notes in Computer Science, Springer.
- [9] Andrew W. Senior, "Recognizing faces in broadcast video," in *IEEE Workshop on Real-Time Analysis and Tracking of Face and Gesture*. 1999, Kerkyra, Greece.
- [10] Phillippe de Cuetos, Chalapathy Neti, and Andrew Senior, "Audio-visual intent to speak detection for human-computer interaction," in *Intl. Conf. On Acoust., Sp., and Sig. Proc.*, Istanbul, Turkey, 2000, IEEE.
- [11] C. Bregler and Y. Konig, "Eigenlips," in *Intl. Conf. On Acoust., Sp., and Sig. Proc.*, Adelaide, Australia, 1994, IEEE.