Visual Subtitles For Internet Videos

Chitralekha Bhat, Imran Ahmed, Vikram Saxena, Sunil Kumar Kopparapu

TCS Innovation Labs - Mumbai, Yantra Park, Thane (West), Maharashtra, INDIA
{bhat.chitralekha, ahmed.imran, vik.saxena, sunilkumar.kopparapu}@tcs.com

Abstract

We present a visual aid for the hearing impaired to enable access to internet videos. The visual tool is in the form of a time synchronized lip movement corresponding to the speech in the video which is embedded in the original internet video. Conventionally, access to the audio or speech, in a video, by the hearing impaired is provided by means of either text subtitles or sign language gestures by an interpreter. The proposed tool would be beneficial, especially in situations where such aids are not readily available or generating such aids is difficult. We have conducted a number experiments to determine the feasibility and usefulness of the proposed visual aid.

Index Terms: Lip movement synthesis, Phone recognition, resource deficient languages

1. Introduction

As per World Health Organization, over 360 million people which account for 5% of the world's total population suffer from hearing loss and a significant majority of them live in developing nations. Moreover, one third of people over the age of 65 years, especially from South Asia, Asia Pacific and Sub-Saharan Africa are affected by disabling hearing loss [1]. A person with hearing impairment, especially acquired deafness in adulthood, can with some training interpret spoken speech by observing lip movements corresponding to the spoken speech.

Lip reading, also known as speech-reading in literature, allows access to speech through visual reading of the movement of the lips, face and tongue in the absence of audible sound. Lip reading also makes use of the information associated with the context, the knowledge of the language, and also the residual hearing of the person [2]. Hearing impairment can prove to be a major handicap especially when a person wishes to understand an internet video while viewing it. Any tool that can make video assessible is useful for the hearing impaired. This motivates our work in developing a tool that allows for viewing a video without having to actually hear the audio track of the video.

Text based subtitles is one way by which a person with hearing loss interprets what is being spoken in a video. However, text subtitles are not always readily available; especially in a country like India where subtitling is not mandated by law unlike in some of the developed nations (example [3, 4]). Moreover, manual generation of subtitles is a long drawn, laborious and an expensive process [5]. An alternative is to automatically generate text subtitles using an Automatic Speech Recognition (ASR) engine, but non-availability of ASR engines for a resource deficient language [6] hinders generation of accurate subtitles, additionally, generating subtitles in the script of the spoken language would be another impediment.

IBM's SiSi (Say It, Sign It) is an automatic sign language generator for spoken audio. SiSi uses a speech recognition module that converts the spoken speech into text; the text is interpreted into gestures, that are used to animate an avatar which signs in British Sign Language [7]. SiSi largely depends on the accuracy of recognition of audio. eSign project was primarily designed to help interpret textual internet content using sign language. eSign synthesizes the signing gestures using Signing Gesture Markup Language (SiGML), along with information regarding speed and viewpoint [8]. However, there are about 200 different sign languages, each with a vocabulary of considerable size. Building an automated system that would generate sign language interpretation of audio would then be complex owing to not only the non-availability of an efficient ASR engine but also the difficulty associated in translation of generated text, to sign language gestures for resource deficient languages.

In this paper, we propose a tool for visual subtitling which is largely based on associating a visual lip movement corresponding to the audio track of the video. The essential idea is based on the fact that recognition accuracies of audio, even for resource deficient languages is higher in viseme space than in the phoneme space. The rest of the paper is organized as follows: We describe the process of generation of visemes in Section 2 and describe the experimental work and evaluation of the proposed tool in Section 3 and conclude in Section 4.

2. Generation of Viseme Sequence

Visual subtitles are essentially a time sequence of visemes corresponding to and in sync with the speech in a given video. Visual subtitles could be in lieu of or in addition to text subtitles, wherein the lip movement for a particular speech will be displayed. It is anticipated that, the user

SLPAT 2013, 4th Workshop on Speech and Language Processing for Assistive Technologies, pages 17–20,
Grenoble, France, 21–22 August, 2013. ©2013 Association for Computational Linguistics

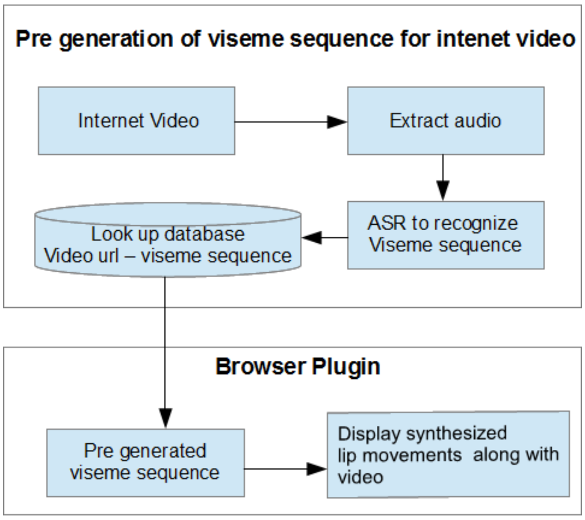
Pre generation of viseme sequence for intenet video
Internet Video
Extract audio
ASR to recognize Viseme sequence
Look up database Video url – viseme sequence
Browser Plugin
Pre generated viseme sequence
Display synthesized lip movements along with video

Figure 1: Overview of Visual Subtitling.

will be able to view the visual subtitles embedded with the original video. Figure 1 represents the proposed the tool. As seen in Figure 1 audio is first extracted from the video. The spoken speech is recognized using a phoneme recognizer and then mapped to the corresponding viseme. Visual subtitles are synthesized using MPEG-4 FAPs for mouth and tongue as defined in [9].

Note 1 *for the purpose of demonstration the videos chosen are predominantly speech audio content.*

2.1. Visemes and MPEG-4 FAPs

Viseme is the basic unit of mouth movement that represents a phoneme or a group of phonemes in the visual domain. We use the standard set of 22 visemes [10]. It is a many phonemes to one viseme mapping with several different phonemes mapped to the same viseme owing to the fact that the lip position for different phonemes is the same; for example the phones /k/ and /g/ correspond to the viseme k or the phonemes /p/, /b/ and /m/ correspond to the viseme p. We first created a mapping between the standard 22 visemes and the Hindi phoneme set as shown in 1 (shows the first 11 visemes only).

Note 2 *The mapping was done so as to include both Hindi and English phonemes to be able to cater to mixed language usage.*

It is desirable to have the lip movement as natural as possible for the user to be able to comfortably understand the audio. MPEG-4 FAPs for mouth and tongue for a given viseme are sufficient to visualize the spoken phoneme completely as can be seen in Figure 2. For each of the 22 visemes, the corresponding FAPs were computed in the form of (x, y) coordinates. For natural visualization of the lip movement, transition between two consequent visemes was simulated by means of a linear interpolation. So in some sense we had intermediate visemes generated.

Table 1: Phoneme to Viseme mapping rule.

Phoneme	Viseme
/sil/	Viseme0
/ae/, /ax/, /ah/, /E/, /EM/, /ai/, /a/	Viseme1
/aa/, /A/, /ah/, /AM/	Viseme2
/ao/, /O/, /au/	Viseme3
/ey/, /eh/, /uh/, /e/, /eh/	Viseme4
/er/, /axr/	Viseme5
/y/, /iy/, /ih/, /ix/, /I/, /IM/, /i/	Viseme6
/w/, /uw/, /U/, /UM, /ux/, /u/, /uh/	Viseme7
/ow/, /o/	Viseme8
/aw/	Viseme9
/oy/	Viseme10
/ay/	Viseme11

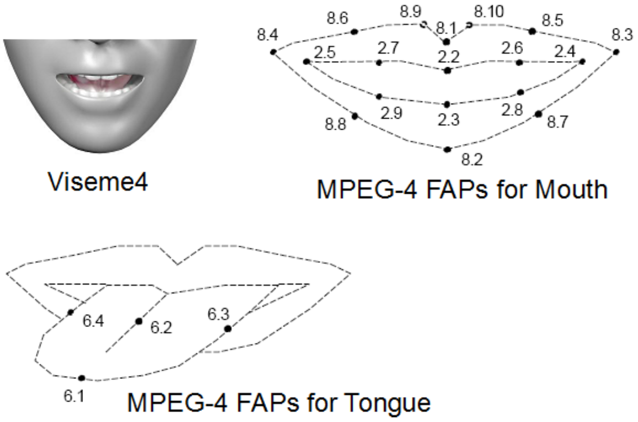

Figure 2: *Viseme and MPEG-4 FAPs for Mouth and Tongue.*

HTK 3.4 ASR [11] was used for phoneme recognition. The recognizer was trained on annotated Hindi data from 100 native speakers of Hindi; each of the speaker spoke 10 sentences each. The HTK 3.4 recognizer performed with 70% correctness on viseme classes when used in the free decoding mode.

Note 3 *The recognition improved by upward of* 10% *for viseme recognition compared to phoneme recognition.*

Manual verification of phone sequence and duration is done to ensure that the lip movement generated by the viseme sequence is a representation of the speech.

A visual subtitle browser plug-in allows the user to view the internet video along with the lip movements corresponding to the speech in the form of a viseme sequence. As stated earlier, given the context, a person with hearing loss would be able to understand the spoken speech from the lip movements, we believe that the video would set the context.

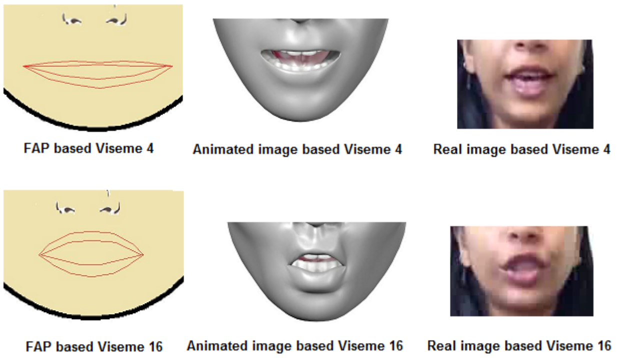
FAP based Viseme 4
Animated image based Viseme 4
Real image based Viseme 4
FAP based Viseme 16
Animated image based Viseme 16
Real image based Viseme 16

Figure 3: A snapshot of the FAP, animated image and real image based visemes.

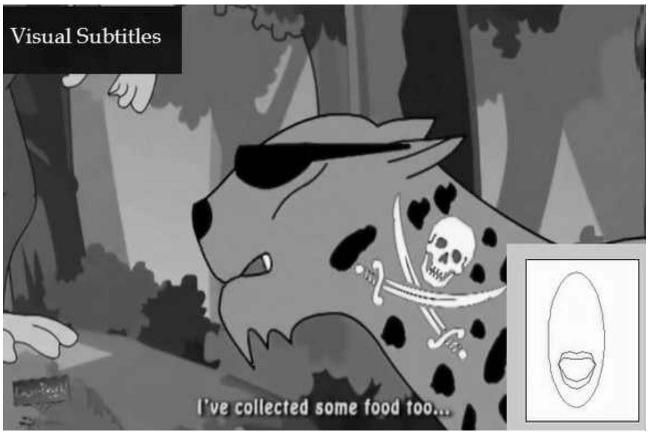

Figure 4: Sample snapshot of a video with visual subtitle.

3. Experimental Results and Discussion

Figure 3. shows the snapshots of FAPs, animated viseme and real viseme that were used to generate the videos. Several videos available on the internet were selected and visual subtitles were generated [12] as mentioned in the earlier sections. While a mix of English and Hindi videos were selected and visual subtitles generated (a snapshot of the video with FAP is shown in Figure 4), the evaluation of the usefulness of the tool was tested on only the English videos because the subjects who evaluated the videos for visual titles were trained to lip read in English.

3.1. Evaluation

All the evaluation results are for English videos only because the subjects were trained in Indian English lip reading. We found access to people who are familiar with lip reading in a language other than English was hard to find. The participants were asked to lip read a video of ten naturally recorded sentences to establish a baseline. Only the mouth portion of the face was used in the baseline videos.

Evaluation was done under different experimental setups, namely,

- Visual subtitles generated using three different visual features, namely, (a) MPEG-4 FAPs, (b) animated viseme images and and (c) real viseme images.
- Visual subtitles with and without the context of video.
- Videos played at different rates, namely, played at their original speed and half the speed.
- Videos comprised of animated clip, classroom lecture, dias/conference lecture.

The participants' understanding based on visual observation of the visual subtitles was evaluated in terms of number of words correctly recognized. In summary, visual subtitles were better understood under the following conditions (a) with the context of video, (b) when played at half the original speed and (c) when generated with real viseme images

Note 4 *MPEG-4 standard does not define FAPs for teeth, which play a significant role in lip reading, hence this aspect needs to be considered during the generation of Visual subtitles using MPEG-4 FAPs.*

4. Conclusions

Visual subtitles are essentially the lip movements corresponding to the audio track in a video. Displaying visual subtitles along with the video would augment the understanding of the content of a video for a person with hearing impairment. Although text subtitles and sign language gesture display can be thought of as alternatives, generating them manually is a tedious task. Automatic generation of text subtitles and sign language gestures for a particular language using ASRs, would require robust ASRs with rich speech corpus. However, lip movements are less language specific, that is one can move from one language to another by modifying the phoneme-visemes mapping. Given these advantages, automatic generation of lip movement from audio emerges as an encouraging solution, especially for resource deficient languages like Hindi. However, automatic generation of lip movements will still be limited by the ASR performance under noisy and with background music. We are also experimenting with other methods like optical flows [13] for generation of transition between visemes.

Acknowledgements: We would like to express our sincere gratitude to all the participants whom we would not like to name. Our thanks are due to Mrs. Alpa Shah for participating and making possible the evaluation of visual subtitles.

5. References

[1] WHO, http://www.who.int/mediacentre/factsheets/fs300/en/, viewed July 2013.

[2] Wikipedia, “Speech reading,” http://en.wikipedia.org/wiki/Speech_reading, viewed July 2013.

[3] CVVA, “U.S. Accessibility Regulations for Online Video Captions,” http://dotsub.com/enterprise/laws, viewed July 2013.

[4] N. K. Aas, “Mandatory subtitling of films for the benefit of the deaf and hard of hearing,” http://merlin.obs.coe.int/iris/2012/1/article34.en.html, viewed July 2013.

[5] Wikipedia, “Subtitle captioning,” http://en.wikipedia.org/wiki/Subtitle_(captioning), viewed July 2013.

[6] I. Ahmed and S. K. Kopparapu, “Speech recognition for resource deficient languages using frugal speech corpus,” in *Signal Processing, Communication and Computing (ICSPCC), 2012 IEEE International Conference on*, 2012, pp. 750–755.

[7] IBM, http://www-03.ibm.com/press/us/en/pressrelease/22316.wss, viewed July 2013.

[8] ISO, “MPEG-4 International Standards.” ISO, Geneva, Switzerland, 1998, no. ISO 14496.

[9] J. R. Kennaway, J. R. W. Glauert, and I. Zwitserlood, “Providing signed content on the internet by synthesized animation,” *ACM Trans. Comput.-Hum. Interact.*, vol. 14, no. 3, Sep. 2007. [Online]. Available: http://doi.acm.org/10.1145/1279700.1279705

[10] Aidreams, “Visemes for character animation,” http://aidreams.co.uk/forum/index.php?page=Visemes-for_Character_Animation, viewed July 2013.

[11] S. J. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book Version 3.4*. Cambridge University Press, 2006.

[12] B. Chitralekha, I. Ahmed, and S. K. Kopparapu, https://sites.google.com/site/awazyp/splat2013, viewed July 2013.

[13] T. A. Faruquie, C. Neti, N. Rajput, L. V. Subramaniam, and A. Verma, “Animating expressive faces to speak in indian languages,” in *National Conference on Communications*, Bombay, 2002, pp. 355–362.