Usable technology for the differently abled-hearing impairment

Chitralekha Bhat and Sunil Kumar Kopparapu*

TCS Innovation Labs – Mumbai,
Tata Consultancy Services Limited,
Yantra Park, Pokhran Road 2,
Thane (West) 400 601, Maharashtra, India
Email: bhat.chitralekha@tcs.com
Email: sunilkumar.kopparapu@tcs.com
*Corresponding author

Abstract: Hearing loss poses a serious challenge to access information that are in the form of audio or video on the internet. Text subtitles or closed captioning is a popular mechanism to make the audio embedded in the videos accessible to the hearing impaired. Alternatively, a sign language interpreter may translate the audio into sign language gestures. However, these aids are either expensive or technologically infeasible. In this paper, we propose a technique to enable a person with hearing impairment *see* the audio. The novel assistive tool is envisioned as a visual aid for the hearing impaired to comprehend the audio contained in a video. The tool produces a speech synchronised visual lip movement sequence for a video. The lip movement sequence is then superimposed onto and displayed along with the original video. A number of experiments have been conducted to determine the feasibility and usefulness of the proposed visual aid.

Keywords: usable technology; assistive technology; hearing impaired; visual audio.

Reference to this paper should be made as follows: Bhat, C. and Kopparapu, S.K. (2018) 'Usable technology for the differently abled-hearing impairment', *Int. J. Humanitarian Technology*, Vol. 1, No. 1, pp.3–18.

Biographical notes: Chitralekha Bhat has been a researcher in the Speech and NLP Group of TCS Innovation Labs, Mumbai since 2011. She has worked at the Digital Audio Processing Lab at IIT Bombay from 2010–2011. She received her MSc in Signal Processing from NTU, Singapore in 2008. She is a music enthusiast and has a training in Carnatic instrument (Veena).

Sunil Kumar Kopparapu received his Doctoral degree in Electrical Engineering from the Indian Institute of Technology, Bombay, India. Before joining TCS Innovation Labs in Mumbai, he was with Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia working on practical image processing and 3D vision problems. In his current role as a Principal Scientist with the TCS Innovations Labs in Mumbai, he is actively working in the areas of speech, script, image and natural language processing

with a focus on building usable systems for mass use in Indian conditions. He has coauthored a book titled *Bayesian Approach to Image Interpretation* and more recently a Springer *Brief on Non-linguistic Analysis of Call Center Conversation* apart from several patents, journal and conference publications.

This paper is a revised and expanded version of a paper entitled 'Visual subtitles for internet videos' presented at 2013 Workshop on Speech and Language Processing for Assistive Technologies, Grenoble, France, 21–22 August 2013.

1 Introduction

Two or three out of every thousand children in the USA are born with a hearing loss in one or both ears (Vohr, 2003) and there are millions in the world who have different degree of hearing loss from mild to profound category, which may affect their performance in day to day living situations (Kacker and Sharma, 2004). Generally, hearing loss greater than 40 decibels (dB) in adults and 30 dB in children, in the better hearing ear, is considered as disabling hearing loss. According to World Health Organization (WHO), about 5% of the world's total population suffer from hearing loss; 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 (WHO, 2015).

Hearing loss is conventionally categorised based on which part of the auditory system is damaged. The three basic types of hearing loss are:

a conductive

b sensorineural

c mixed hearing loss.

Cochlear implants, hearing assistive technologies, audiologic rehabilitation, hearing aids, etc., are some means that help in integrating people with hearing impairment into society (ASHA, 2003). Cochlear implants are electronic devices that are surgically inserted to provide a heightened sense of sound for people suffering from sensorineural hearing loss, however this does not help restore normal hearing; subsequently people with cochlear implants have to depend on some form of external aid. Any assistive technology (for example personal FM systems, TV listening devices, or ability to lip read) become a dire necessity. Lip reading, also known as speech-reading allows access to speech audio through visual reading of the movement of the lips, face and tongue in the absence of audible sound. Lip reading is more effective when it also makes use of the information associated with the context of the speech, the knowledge of the language (Wikipedia, 2003). Hearing impairment can prove to be a major handicap especially when a person wishes to understand a video lecture (popularised by online courses) while viewing it. Although guidelines have been laid out to make internet content accessible to the disabled (CVVA, 2012; Aas, 2012), the mandates are mostly

applicable to Government websites; leaving room to address accessibility issues in non-Government internet content as well as non-public-non-internet content.

With increasing digitisation, massive open online courses (MOOCs) have become very popular (Abeer and Miri, 2014) and with these learning has primarily turned into one of learning from video lectures (Ronchetti, 2010). While this has brought in a paradigm shift from classroom learning, it has made it difficult for the hearing impaired because the audio content of the videos are not accessible for the hearing impaired.

Subtitling or closed captioning is by and large the most widely used tool to represent audio to a person with hearing loss especially in the television broadcasting. In closed captioning, the text corresponding to the audio being spoken in the video appears time synchronised and superimposed on the video. A person with hearing impairment can access what is being spoken in a video by reading the text. While subtitling is mandated in several countries, for example, CVVA (2012) and Aas (2012), unfortunately in most developing countries, like India, this is not mandated and as a result, text subtitles are not always readily available for most of the videos.

While manual generation of subtitles (text corresponding to the audio plus alignment with the audio) is a long drawn, labourious and an expensive process (Wikipedia, 2004), an alternative is to automatically generate text using a large vocabulary automatic speech recognition (ASR) engine. While there have been huge strides in the area of speech recognition, especially with the introduction of deep learning (Deng and Yu, 2014; Weng et al., 2014) the development has been essentially for resource rich languages like English (Kopparapu, 2014). However, non-availability of good ASR engines for resource deficient languages (Ahmed and Kopparapu, 2012) hinders automatic generation of accurate subtitles, additionally, generating subtitles in the script of the spoken language is yet another impediment without the availability of a robust and extensive sound-glyph parallel corpus (Bhat and Kopparapu, 2014). IBM's Say It, Sign It (SiSi) converts audio into sign language instead of text. Like subtitling, SiSi uses speech recognition to convert the spoken speech into text; the text is used to animate an avatar which signs in British Sign Language (IBM, 2007). The performance of SiSi largely depends on the accuracy of recognition of audio.

Extensive literature exists which discusses synthesising lip movement from either speech audio or from text. In Benoît and Goff (1998), authors discuss several techniques for developing a visual speech synthesiser for French language including the use of 2D and 3D parametric models which cover the main components of a human face. The lip movement sequence is generated from the audio using hidden Markov models in Yamamoto et al. (1998); however the accuracy of the constructed lip movement sequence relies heavily on the underlying ASR which is not yet mature for resource deficient languages like Hindi and other Indian languages. In Lewis and Parke (1986), authors discuss the lip-sync synthesis for character animation using a parametric model for the human face. Yet another technique called 'Video Rewrite', uses existing video footage to automatically create a new video of a person mouthing words that she did not speak in the original footage (Bregler et al., 1997). This technique is of use only when there is large amount of video footage, concentrating on the face of a single speaker (like a news reader). In Massaro et al. (2007), authors outline a method of modelling speech distinctions within computer-animated talking heads that utilise the manipulation of speech production articulators for selected speech segments. In Hoon et al. (2015), a usable, automated digital speech system for lip-sync animation was

developed. Synchronisation tricks were used to improve accuracy and create realistic visual impression.

Similar to what we propose in this paper, some methods have been proposed in literature, which try to overcome the inaccuracies of an ASR engine. For example, Lavagetto (1995) discusses the possibility of making use of graphic animation of telephonic speech for easing telephonic communication for the hearing impaired. Also, Rathinavelu et al. (2006) is directed towards building a tool for hearing impaired, wherein data for lip-sync model was collected from video image and magnetic resonance imaging (MRI) techniques. The articulatory movements of the lip sync model were presented along with virtual reality (VR) objects in an interactive multimedia (IMM) interface. This IMM interface was used to teach small vocabulary to hearing impaired (HI) children.

In this paper, an extension of Bhat et al. (2013a) with more granular details, we propose a novel and usable visual subtitling technique to make the audio content of a video accessible to the hearing impaired, which is the main contribution of this paper. The essential idea is to convert the audio in a video into a phoneme sequence and then display the corresponding viseme for the duration of the phoneme. The main advantage of this process (phoneme to viseme) is that though the phoneme recognition accuracy is poor, the confusing phonemes generally map into the same viseme, thereby providing robustness. The robustness comes due to 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 organised as follows: we describe the process of generation of viseme sequence for a given audio in Section 2 and describe the experimental work and evaluation of the proposed tool in Section 3; we conclude in Section 4.

2 Visual subtitling

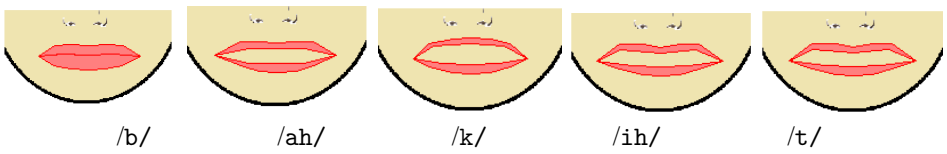
Phoneme is the smallest unit of sound and any word we speak can be represented as a sequence of concatenated phonemes. A viseme is the visual representation of a phoneme, for example, Figure 1 shows the viseme corresponding to the phoneme /r/.

Figure 1 Viseme corresponding to the phoneme /r/ (see online version for colours)

For example the English word 'bucket' is spoken as a sequence of connected phonemes, namely, /b/ /ah/ /k/ /ih/ /t/. The corresponding viseme sequence is shown in Figure 2.

Visual subtitles are essentially a time sequence of visemes corresponding to and in time synchronisation with the speech in a given video. Figure 3 captures the entire process of constructing a viseme sequence of an audio signal for the utterance 'good morning'.

Figure 2 Viseme sequence corresponding to the word `'bucket'` (≡ /b/ /ah/ /k/ /ih/ /t/) (see online version for colours)



As seen in Figure 3, the audio track is extracted from the video. The phoneme sequence is identified using a phone recognition engine (Young and Young, 1994). The phone recognition engine not only identifies the phonemes in the audio stream but also gives the duration (length) of each of the phoneme in the audio (see Table 1). Using a phoneme to viseme look-up table (see Table 2 and Figure 4) and the duration information for each phoneme, lip movement sequence is generated. This video of the lip movement sequence, consisting of the visemes shown in Figure 4, is superimposed onto the original video to be played along with it. The major steps in creating visual subtitles for a given video are

1 process video to extract the audio

2 recognition phonemes in the audio to identify

   - phoneme sequence
   - phoneme duration

3 pick the viseme corresponding to the phoneme

4 build a natural viseme sequence or visual subtitles corresponding to the phoneme sequence

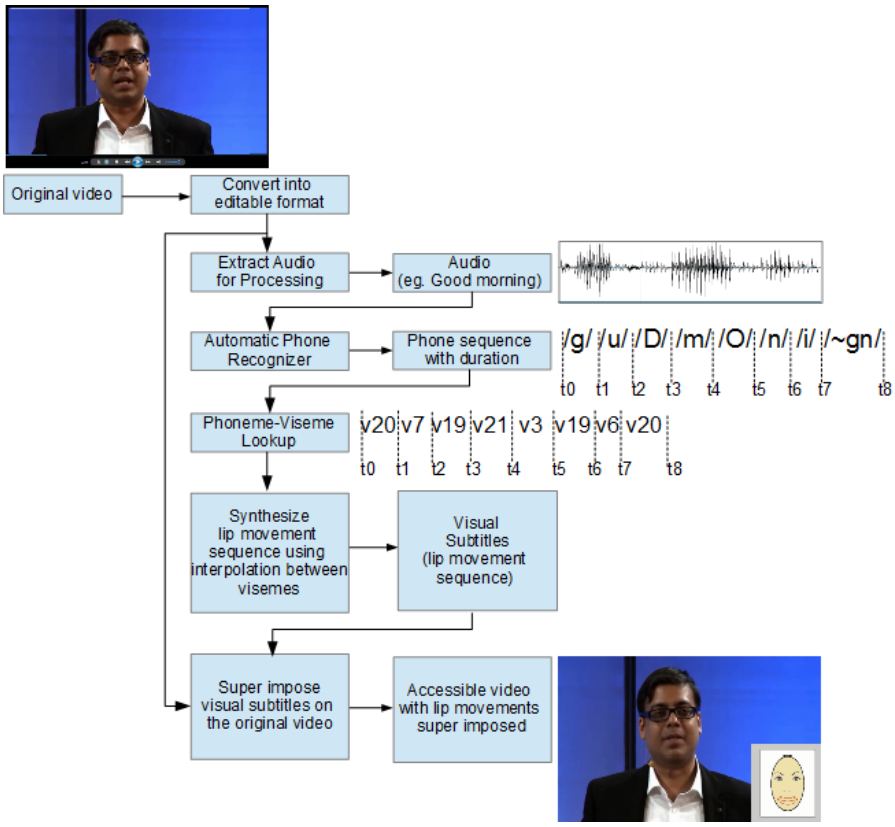
5 superimpose this visual subtitles onto the original video.

We elaborate and delve into the details of the construction of lip movement sequence in the following sections.

2.1 Video processing to extract audio

Video is processed to extract audio from the video by using a well known library called `ffmpeg` (2016). Audio file to be recognised by the phoneme recogniser needs to meet the requirement of the phoneme recognition engine in terms of the sampling rate and format so as to be compatible with the training data used to build the acoustic models. We make sure the audio is sampled at 16 kHz and is in wave format.

Figure 3 Overview of visual subtitling (see online version for colours)



Source: The image shown is a snapshot from the video in Tekriwal (2013)

2.2 *Phoneme recognition module*

HTK 3.4 toolkit (Young et al., 2006), an ASR engine framework, was used for phoneme recognition. The recogniser acoustic models were built by training the engine using annotated Hindi data from 100 native speakers of Hindi; each of the speaker spoke 10 sentences each. In the phoneme recognition mode, the ASR is configured to recognise the phoneme sequence in an utterance by using acoustic model score alone (no language model was used). This makes it possible for cross language recognition and renders the process fairly independent of language.

The output of the phone recogniser is the recognised phone and the duration of the phone in msec (milli seconds). For example, the spoken utterance `'bucket'` could result in Table 1. This forms the input to the viseme synthesiser.

2.3 *Lip movement sequence synthesis*

By definition, a viseme is the visual equivalent of a phoneme in a spoken language, hence a viseme can be considered to represent a phoneme or a group of phonemes in

the visual domain. For example, the viseme 18 (Table 2) is common for the phonemes /f/ and /v/ and similarly /k/ and /g/ have the same visual representation. We use the standard set of 22 visemes (Aidreams, 2013) (also see Figure 4). Though the phoneme accuracy of a phone recognition engine is poor, it is the *many phonemes to one viseme mapping*, namely, several different phonemes mapped to the same viseme that makes the process of visual subtitling robust.

We first created a mapping between the standard 22 visemes and the phonemes from both Hindi and English language as shown in Table 2. The many to one mapping was constructed so as to include both Hindi and English phonemes to be able to cater to mixed language usage which is prominent in the Indian sub-continent (Bhuvanagiri and Kopparapu, 2010). MPEG-4 facial animation parameters (FAP) for mouth and tongue for a given viseme are sufficient to visualise the spoken phoneme completely (Pockaj and Lavagetto, 2001) as seen in Figure 5.

Table 1 Output of a phoneme recogniser (for the spoken word 'bucket')

Phoneme	*Duration*
/b/	t_1
/ah/	t_2
/k/	t_3
/ih/	t_4
/t/	t_5

Table 2 Phoneme to viseme mapping

Viseme no.	*Phoneme*
0	/sil/
1	/Ae/ /ax/ /ah/ /E/ /ae/ /EM/ /ai/ /a/ /aM/
2	/Aa/ /A/ /AM/
3	/ao/ /O/ /au/
4	/ey/ /eh/ /uh/ /e/ /eM/ /ey/
5	/er/ /axr/
6	/y/ /iy/ /ih/ /ix/ /i/ /I/ /IM/ /iM/
7	/w/ /uw/ /U/ /UM/ /ux/ /u/ /uh/
8	/ow/ /o/
9	/aw/
10	/oy/
11	/ay/
12	/h/
13	/r/
14	/l/ /el/
15	/s/ /z/
16	/sh/ /ch/ /jh/ /zh/ /j/ /S/
17	/th/ /dh/ /Th/ /Dh/
18	/f/ /v/
19	/d/ /t/ /n/ /ta/ /da/ /T/ /D/ /nn/ /N/ /nx/
20	/k/ /g/ /ng/ /eng/ /˜gn/ /˜jn/ /en/
21	/p/ /b/ /m/ /ph/ /mm/ /em/

Figure 4 Complete list of visemes (see online version for colours)

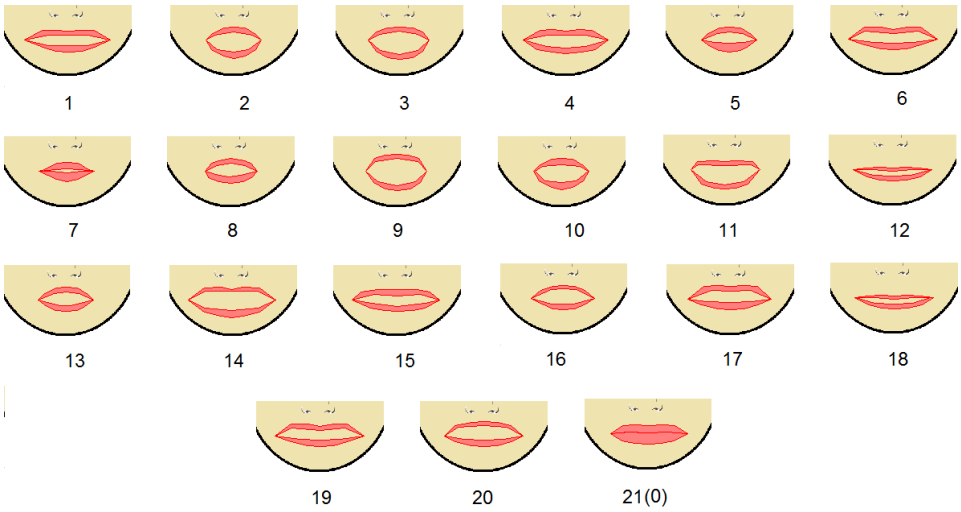
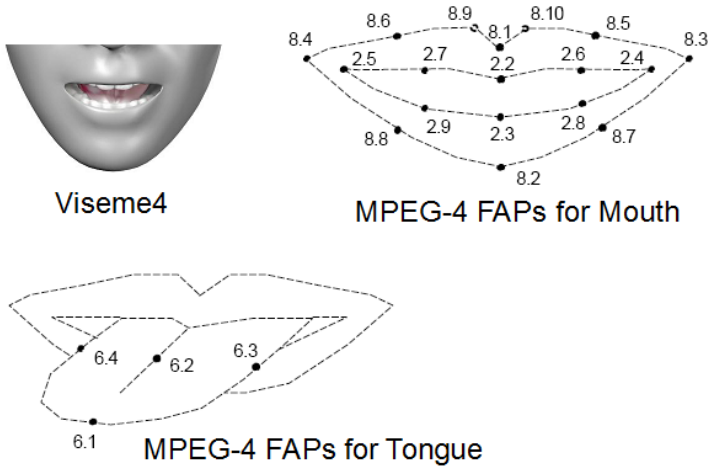


Figure 5 A typical viseme and its corresponding MPEG-4 FAPs for mouth and tongue (see online version for colours)



For each viseme there were in all 18 (x, y) coordinates corresponding to the lips $(2.2, 2.3, \cdots 2.9$ and $8.1, 8.2 \cdots, 8.10$ in Figure 5) and 4 points corresponding to the tongue $(6.1, 6.2, 6.3, 6.4$ in Figure 5); so each viseme was represented by $\{(x_i, y_i)\}_{i=1}^{22}$ points.

The (x, y) coordinates of the viseme corresponding to the identified phoneme (say /ah/) is selected and displayed for the corresponding duration namely, t_2 msec (see Table 1), before being replaced by the viseme corresponding to the next phoneme namely, /k/. However this transition from one phoneme to the next produces a visible distinct jump in the viseme sequence (/ah/ $\rightarrow$ /k/). For continuous and smooth visual transition from one viseme to another, it is necessary that the synthesised lip movements be natural. For natural visualisation of the lip movement, transition between two consequent visemes was smoothened using a linear interpolation technique.

Let each viseme be represented by $N = 22$ FAPs. Let the coordinates for the first viseme (V_1) be

$$V_1 = (x_i^1, y_i^1) \text{ for } i = 1, 2, \cdots, N$$

and the coordinates for the viseme following V_1 (say V_2) be represented as

$$V_2 = (x_i^2, y_i^2) \text{ for } i = 1, 2, \cdots, N$$

We build L intermediate FAP states between V_1 and V_2 by way of smooth linear interpolation, namely,

$$V_{1\to 2}^j = (1 - c_j) * (V_1) + c_j * (V_2) \text{ where } c_j = \frac{j}{L} \text{ and } j = 1, 2, \cdots L. \quad (1)$$

Equation (1) generates L intermediate viseme coordinates which makes the visual transition from $V_1 \to V_2$ smooth and more natural. An example of the phoneme-viseme chart for the speech utterance `'good morning'` can be seen in Table 3.

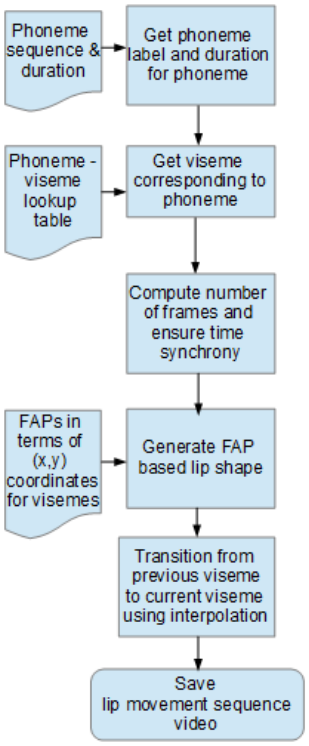
Table 3 Phoneme-viseme chart for speech utterance `'good morning'`

```
<phoneme-viseme>
    <word#>good</word#>
            <phoneme>/g/</phoneme>
                <viseme>20</viseme>
                <duration>12</duration>
            <phoneme>/u/</phoneme>
                <viseme>7</viseme>
                <duration>20</duration>
            <phoneme>/D/</phoneme>
                <viseme>17</viseme>
                <duration>10</duration>
    <word#>morning</word#>
            <phoneme>/m/</phoneme>
                <viseme>21</viseme>
                <duration>18</duration>
            <phoneme>/O/</phoneme>
                <viseme>3</viseme>
                <duration>28</duration>
            <phoneme>/n/</phoneme>
                <viseme>19</viseme>
                <duration>15</duration>
            <phoneme>/i/</phoneme>
                <viseme>6</viseme>
                <duration>20</duration>
            <phoneme>/~ng/</phoneme>
                <viseme>20</viseme>
                <duration>12</duration>
</phoneme-viseme>
```

2.4 Superimposing the visual subtitles into the original video

The visual subtitles or the lip movement sequence video thus generated was superimposed onto the original video to create the visually subtitled video. The time alignment of the visual subtitles to the original video was ensured by using the same frame-rate for generation of visual subtitles as that of the original video. Figure 6 depicts the algorithm for lip movement sequence generation.

Figure 6 Algorithm for lip movement sequence generation (see online version for colours)



3 Experimental results and discussion

We describe in detail the experiments we conducted to verify the usability of the proposed tool. Importantly, we conducted a number of experiments with participants who were hearing impaired and really in need of such a tool.

3.1 Experimental setup

- *Demography of the participants:* We had a total of five participants (two males and three females). The youngest participant was ten years old, while the oldest participant was 22 years old. All of them were profoundly deaf with about only 5%–10% hearing left. All of them were attending or had attended mainstream school; one of the participants was well versed in sign language All of them

enjoyed watching videos on the Internet or television, but stated that they could not watch much because not many videos had text subtitles. They also mentioned that they required the assistance of an interpreter while watching some videos.

Note that, all the experimental analysis reported in this paper is based on the response from these participants.

- *Selection of videos:* Videos were selected such that

  a the audio component was predominantly speech (video tutorials)

  b to include classroom lectures, dias conferences, etc., where the speaker's mouth is not always visible in the video.

  We assume that the context is provided by the visuals in the video, thereby assuming that a person with hearing loss could comprehend the spoken speech by following just the lip movements. Several videos available on the internet were selected and visual subtitles were generated (Bhat et al., 2013b) as described earlier. A mix of English, Hindi and a few regional Indian language videos were selected and visual subtitles generated (a snapshot of the video with FAP is shown in Figure 8).

- *Baseline establishment:* This step is important to understand the speech or lip reading capabilities of the participants. Evaluation results are reported for English videos only, since the participants were trained in Indian English lip reading. The participants were asked to lip read a video of ten naturally recorded sentences to establish a baseline of the lip reading ability of the participants. Only the mouth portion of the face was used in the baseline videos (see Figure 7, column 3). As expected, each participant fared differently; and the participant with best lip reading ability understood the baseline sentences with a word error rate (WER) of 4%, where WER is defined as

$$\text{WER} = \left(1 - \frac{\#\text{ of correctly recognised words}}{\#\text{ of words present in the audio}}\right) \times 100 \tag{2}$$

- *Types of videos:* Five original videos were used to create eight different types of subtitled videos as shown in Table 4. Context of what is being spoken can be gathered from the video content. For example, the participant may lip read a word as `'park'` or `'bark'` since both /p/ and /b/ have the same viseme. However, if the video is showing a dog, the spoken word is most likely `'bark'`, the video depicting a dog thus provides the context that will help in resolving between the two choices `'park'` and `'bark'`.

- *Viewing setup:* The subtitled videos were played to the participants in a group. The above mentioned videos were displayed to participants in random order to avoid participant bias as well to enable us to understand the contribution of each aspect such as context, type of visual subtitling (animated image, MPEG-4 FAP, real image) and speed. Each participant was asked to write down what they understood separately on a piece of paper without interacting with each other. There were no playbacks done, each video was displayed only once.

Table 4 Types of visual subtitled videos

S. no.	*Type (see Figure 7)*	*Context*	*Speed*
1	Animated visemes	No	Original
2	Animated visemes	Yes	Original
3	Animated visemes	No	Half original speed
4	Animated visemes	Yes	Half original speed
5	MPEG-4 FAP	No	Original
6	MPEG-4 FAP	Yes	Original
7	MPEG-4 FAP	No	Half original speed
8	MPEG-4 FAP	Yes	Half original speed
9	Real image	No	Original
10	Real image	Yes	Original
11	Real image	No	Half original speed
12	Real image	Yes	Half original speed

Figure 7 A snapshot of the FAP (column 1), animated image (column 2) and real image-based (column 3) visemes to represent viseme 4 and 16 (see online version for colours)

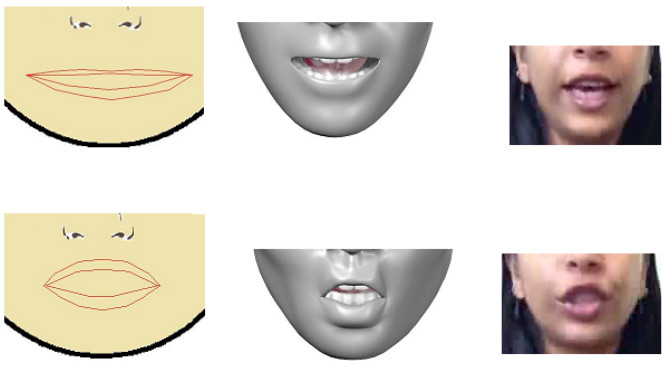


Figure 8 Sample snapshot of a video with visual subtitle, (a) video frame where speaker is visible (b) video frame where focus is not on the speaker (see online version for colours)

Source: Tekriwal (2013)

3.2 Evaluation

It is desirable to have a quantitative evaluation in terms of number of words correct or WER (2). However, this requires extensive participation and several iterations for the participants to be able to provide such a feedback. Hence, we restricted our study to understand the value addition that the visual subtitles may provide to the hearing impaired. Here we present a qualitative evaluation based on our observations.

The videos were shown in random order to avoid any bias. The participants' understanding of the visual subtitles was evaluated based on WER (2). A qualitative analysis was done based on the participant experience and feedback. Participant feedback and our inferences are as below.

Observation 1: All the participants unanimously agreed that the visual subtitles $1, 3, 5, 7, 9$ and 11, were difficult to understand and consequently they were unable to identify most words.

> These were visual subtitles where no video context was provided. Participants seem to depend heavily on the context for lip reading.

Observation 2: Participants were unable to understand the audio content projected by visual subtitles $1, 2, 3, 4$ despite visual subtitles 2 and 4 providing video context.

> These visual subtitles were created using animated visemes. There was no natural or smooth transition between images which made the viewing experience uncomfortable due to a jerky transition between visemes. Additionally, these animated visemes correspond to American English phoneme-viseme mapping. Contribution of variation in accents (wherein the participants are trained to lip read Indian English and the visemes correspond to American English) played a role in deteriorated understanding.An extensive analysis using an Indian and a Croatian speaker (Bhat and Kopparapu, 2015), showed that participants trained to lip read Indian English found it easier to understand the content (through lip reading) of an Indian English speaker's speech.

Observation 3: Participants' experience of viewing visual subtitles $5, 6$ was better in terms of smooth transitioning of lip movements. However,they were unable to clearly understand the words spoken. Providing the context and playing at half speed as done in 8 did help increase the understandability, wherein each participant recognised the words drawing from their experience.

> There was a strong correlation observed where the participant who fared well in the baseline test, fared best here as well. Based on the confusions in words identified by the participants, it was observed that tongue and teeth are essential cues for lip reading. This was confirmed by the participants as well.

Observation 4: For visual subtitles $9, 10, 11, 12$, participants were able to understand the spoken words more clearly when the videos were played at half the speed and performed better than all the eight different types of visual subtitles viewed earlier.

> These were visual subtitles constructed using visemes from real images of an Indian English speaker. These visual subtitle videos were jerky owing to the lack of smooth transition between visemes.

Based on the above observations, we can conclude that the key factors to increased understanding would be

a constructing the visual subtitles using Indian English visemes,

b provide a smooth transition

c to ensure teeth and tongue for each viseme are visible

d provide video context.

Going forward we plan to include the FAPs for tongue (see Figure 5) and explore mechanisms of extending FAPs to include teeth; the additional MPEG-4 FAP based viseme sequence or lip movement sequence should enhance our visual subtitling tool significantly.

4 Conclusions

With ease in digitisation, there has been an increase in the generation and distribution of video data. While most of the video generated is for entertainment. There is a significant increase in the amount of video data in the infotainment space. As a consequence, there has been a spurt in the generation of video lectures mostly driven by MOOCs. The availability of video lectures has brought in a paradigm shift from the conventional classroom learning; however it has made it difficult for the hearing impaired because most often, the audio that is a crucial aspect of these videos is not accessible.

Considering the limitations of the available tools for the hearing impaired in terms of instruments that would make sounds audible like cochlear implants, hearing aids, etc., assistive technologies such as lip reading play an important role in assimilation of information. Though text subtitles and sign language gesture display can be thought of as tools to make video accessible; manual generation of text subtitles is a tedious task. Automatic generation of text subtitles and sign language gestures for a particular language using ASRs, requires robust ASRs in that language, which in turn demands a rich speech corpus. Such speech corpora are unavailable for most Indian languages. Visual subtitles, as described in this paper, is essentially an automatic tool to make videos accessible to the hearing impaired. The idea is to generate a lip movement sequence corresponding to the audio track in a video and displaying visual subtitles (as a picture in picture – see Figure 8) in the video. This will enable a person with hearing impairment to comprehend the content of a video better. Additionally, lip movements are loosely coupled to a language and can be generated with ease using phoneme-viseme mapping as compared to text subtitles. Moreover, *many phonemes to one viseme mapping*, namely, several different phonemes mapped to the same viseme makes the visual subtitling robust. Given these advantages, automatic generation of lip movement from audio emerges as an encouraging solution, especially for resource deficient languages (most Indian languages).

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, a lip reading trainer, for participating and making possible the evaluation of visual subtitles. The authors would also like to thank the members of the TCS Innovation Labs – Mumbai.

References

Aas, N.K. (2012) 'Mandatory subtitling of films for the benefit of the deaf and hard of hearing' [online] http://merlin.obs.coe.int/iris/2012/1/article34.en.html (accessed March 2016).

Abeer, W. and Miri, B. (2014) 'Students' preferences and views about learning in a MOOC', *Procedia – Social and Behavioral Sciences, ERPA, International Congress on Education*, Istanbul, Turkey, 6–8 June 2014, Vol. 152, No. 0, pp.318–323.

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

Aidreams (2013) 'Visemes for character animation' [online] http://aidreams.co.uk/forum/index.php?page=Visemes-for_Character_Animat% ion (accessed March 2016).

ASHA (2003) [online] http://www.asha.org/public/hearing/Treatment/ (accessed March 2016).

Benoît, C. and Goff, B.L. (1998) 'Audio-visual speech synthesis from french text: eight years of models, designs and evaluation at the ICP', *Speech Communication*, Vol. 26, Nos. 1–2, pp.117–129.

Bhat, C. and Kopparapu, S. (2014) 'Constructing a sound-glyph database for subtitling videos', in *Oriental COCOSDA*, Phuket, Thailand, September.

Bhat, C. and Kopparapu, S.K. (2015) 'Viseme comparison based on phonetic cues for varying speech accents', in *INTERSPEECH, 16th Annual Conference of the International Speech Communication Association*, Dresden, Germany, September 6–10, pp.3412–3416.

Bhat, C., Ahmed, I., Saxena, V. and Kopparapu, S.K. (2013a) 'Visual subtitles for internet videos', in *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*, Association for Computational Linguistics, Grenoble, France, August, pp.17–20.

Bhat, C., Ahmed, I. and Kopparapu, S.K. (2013b) [online] https://sites.google.com/site/awazyp/splat2013 (accessed March 2016).

Bhuvanagiri, K. and Kopparapu, S. (2010) 'An approach to mixed language automatic speech recognition', *Oriental COCOSDA*, Kathmandu, Nepal.

Bregler, C., Covell, M. and Slaney, M. (1997) 'Video rewrite: driving visual speech with audio', in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH*, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, pp.353–360.

CVVA (2012) 'U.S. accessibility regulations for online video captions' [online] http://dotsub.com/enterprise/laws (accessed March 2016).

Deng, L. and Yu, D. (2014) 'Deep learning: methods and applications', *Foundations and Trends in Signal Processing*, Vol. 7, Nos. 3–4, pp.197–387.

ffmpeg (2002) [online] http://www.ffmpeg.org/ (accessed March 2016).

Hoon, L.N., Rahman, K.A.A.A. and Chai, W.Y. (2015) 'Framework development of real-time lip sync animation on viseme based human speech', *Jurnal Teknologi*, Vol. 75, No. 4.

IBM (2007) [online] http://www-03.ibm.com/press/us/en/pressrelease/22316.wss (accessed March 2016).

Kacker, S. and Sharma, R. (2004) 'People supporting the hearing impaired', in J-I. Suzuki, T. Kobayashi and K. Koga (Eds.): *Hearing Impairment*, pp.425–432, Springer, Japan.

Kopparapu, S. (2014) *Non-Linguistic Analysis of Call Center Conversations*, SpringerBriefs in Electrical and Computer Engineering, Springer International Publishing.

Lavagetto, F. (1995) 'Converting speech into lip movements: a multimedia telephone for hard of hearing people', *Rehabilitation Engineering, IEEE Transactions on*, March, Vol. 3, pp.90–102.

Lewis, J.P. and Parke, F.I. (1986) 'Automated lip-synch and speech synthesis for character animation', *SIGCHI Bull.*, May, Vol. 17, pp.143–147.

Massaro, D., Cohen, M. and Beskow, J. (2007) *Visual Display Methods for in Computer-animated Speech Production Models*, May 29, US Patent 7,225,129.

Pockaj, R. and Lavagetto, F. (2001) 'An efficient use of mpeg-4 FAP interpolation for facial animation at 70 bits/frame', *IEEE Trans. Circuits Syst. Video Techn.*, Vol. 11, No. 10, pp.1085–1097.

Rathinavelu, A., Thiagarajan, H. and Savithri, S. (2006) 'Evaluation of a computer aided 3d lip sync instructional model using virtual reality objects', in *Proc. 6th Intl Conf. Disability*, Virtual Reality & Assoc. Tech., Esbjerg, Denmark.

Ronchetti, M. (2010) 'Using video lectures to make teaching more interactive', *iJET*, Vol. 5, No. 1, pp.45–48.

Tekriwal, G. (2013) 'The magic of vedic math' [online] https://www.youtube.com/watch?v=grkWGeqW99c (accessed March 2016).

Vohr, B. (2003) 'Overview: infants and children with hearing loss-part I', *Ment. Retard Dev. Disabil. Res.*, Vol. 9, pp.62–64.

Weng, C., Yu, D., Watanabe, S. and Juang, B.F. (2014) 'Recurrent deep neural networks for robust speech recognition', in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, Florence, Italy, May 4–9, pp.5532–5536.

WHO (2015) [online] http://www.who.int/mediacentre/factsheets/fs300/en/ (accessed March 2016).

Wikipedia (2003) 'Speech reading' [online] http://en.wikipedia.org/wiki/Speech_reading (accessed March 2016).

Wikipedia (2004) 'Subtitle captioning' [online] http://en.wikipedia.org/wiki/Subtitle_(captioning) (accessed March 2016).

Yamamoto, E., Nakamura, S. and Shikano, K. (1998) 'Lip movement synthesis from speech based on hidden Markov models', in *Automatic Face and Gesture Recognition, Proceedings, Third IEEE International Conference on*, April, pp.154–159.

Young, S. and Young, S. (1994) *The HTK Hidden Markov Model Toolkit: Design and Philosophy*, Entropic Cambridge Research Laboratory, Ltd., Cambridge, Vol. 2, pp.2–44.

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