Dynamic Captioning: Video Accessibility Enhancement for Hearing Impairment

Richang Hong, Meng Wang, Mengdi Xu† Shuicheng Yan† and Tat-Seng Chua
School of Computing, National University of Singapore, 117417, Singapore
†Department of ECE, National University of Singapore
{dcsrh, eleyans, chuats}@nus.edu.sg, {eric.mengwang, mengdi.xu}@gmai.com

ABSTRACT

There are more than 66 million people suffering from hearing impairment and this disability brings them difficulty in video content understanding due to the loss of audio information. If the scripts are available, captioning technology can help them in a certain degree by synchronously illustrating the scripts during the playing of videos. However, we show that the existing captioning techniques are far from satisfactory in assisting the hearing impaired audience to enjoy videos. In this paper, we introduce a scheme to enhance video accessibility using a *Dynamic Captioning* approach, which explores a rich set of technologies including face detection and recognition, visual saliency analysis, text-speech alignment, *etc.* Different from the existing methods that are categorized as static captioning, dynamic captioning puts scripts at suitable positions to help hearing impaired audience better recognize the speaking characters. In addition, it progressively highlights the scripts word-by-word via aligning them with the speech signal and illustrates the variation of voice volume. In this way, the special audience can better track the scripts and perceive the moods that are conveyed by the variation of volume. We implemented the technology on 20 video clips and conducted an in-depth study with 60 real hearing impaired users. The results demonstrated the effectiveness and usefulness of the video accessibility enhancement scheme.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems-Evaluation/methodology; C.4 [**Performance of Systems**]: Design studies; H.1.2 [**Models and Principles**]: User/Machine Systems-Human factors

General Terms

Experimentation, Human Factors, Performance

Keywords

Accessibility, Dynamic Captioning, Hearing Impairment

MM'10, October 25–29, 2010, Firenze, Italy.

1. INTRODUCTION

Video is an important information carrier that presents visual and audio content in live form. With rapid advances of capturing and storage devices, networks and compression techniques, videos are growing in an explosive rate and play an increasing important role in peoples' daily life. However, there are millions of people that are suffering from hearing impairment. They are fully or partially unable to perceive sound. It is estimated that there are more than 66 million people with hearing impairment, of which about 41% cannot hear any speech at all and 59% are able to hear only if words are shouted around their ears [1]. This disability brings them great difficulty in comprehending video content as audio information is lost.

There are two typical approaches to helping these special audience better access videos. The first is "*direct access*", which provides access as part of the previously developed system [2]. However, merely providing access is not sufficient for hearing impaired audience. Therefore, most attempts have focused on the second approach, i.e., the so-called assistive approach. For a large family of videos that have associated scripts[1] , such as movies, television programs and documentary, captioning is the most widely-applied assistive technique. By synchronously illustrating the scripts during the playing of videos, hearing impaired audience can obtain the necessary information from texts.

Generally, captioning can be categorized into open captions and closed captions according to whether it is able to be activated by users; and it can also be presented in a variety of styles (several examples can be found in Fig. 1). However, the existing captioning methods most resemble each other as the scripts are simply demonstrated in a fixed region and they are illustrated statically. Although hearing impaired audience can get certain information from the scripts, they still encounter difficulty in the following aspects:

(1) Confusion on the speaking characters. When multiple characters are involved in a scene, hearing impaired audience need to judge from which person the scripts come, and this adds their difficulty of content understanding and also degrades their experience of video enjoyment.

[1]These videos are also called *multimedia videos*. Although there are also many videos that have no script information, as mentioned in Section 5, our scheme can be extended to deal with general videos by further exploring speech recognition and speaker identification technologies. Actually this work is just our primary step towards helping hearing impaired people better access video content.

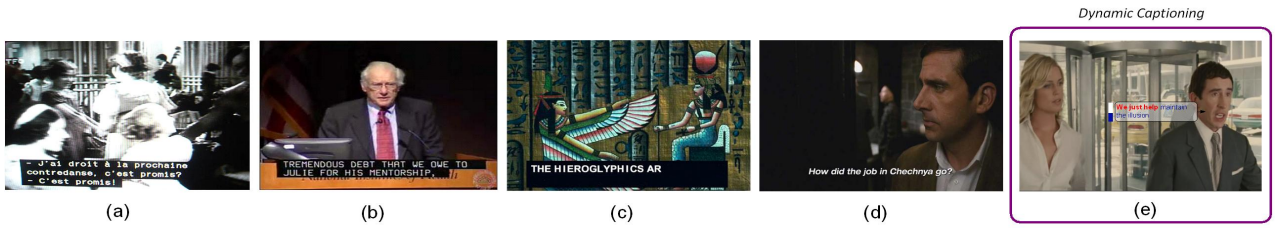


Figure 1: Examples of different captioning styles: (a) scroll-up captioning; (b) pop-up captioning; (c) pain-on captioning; (d) cinematic captioning; and (e) dynamic captioning. The first four techniques can be categorized as static captioning, and different from them, dynamic captioning in (e) benefits hearing impaired audience by presenting scripts in suitable regions, synchronously highlighting them word-by-word and illustrating the variation of voice volume.

(2) The tracking of captioning. In video playing, there is no hint on the duration of each piece of script. As speaking pace can vary significantly, the duration of the script presentation will also vary over a wide range. This brings hearing impaired audience difficulty in the tracking of scripts. For example, they may miss a part of a sentence when the character is speaking rapidly.

(3) The lost of volume information. The variation of volume conveys important information about the emotion [27] [28]. For example, the sound of a character will be loud if he/she becomes happy or angry. However, such information is lost in the existing captioning technology.

Therefore, the existing captioning approach is far from satisfactory in assisting hearing impaired audience. A recent study reporting that the conventional captioning approach can hardly add significant information for hearing impaired audience's perception [4][5]. One major reason is that the audience can hardly track the scripts and match them with visual content rapidly.

In this work, we propose a novel approach named *dynamic captioning* to enhance the accessibility of videos for hearing impairment. Compared with the existing captioning methods which are categorized as static captioning, dynamic captioning is able to help hearing impaired users match the scripts with the corresponding characters. Dynamic captioning is also able to synchronize the scripts word-by-word with the speech as well as highlight the variation of voice volume. In this way, the aforementioned three problems can be addressed. Figure 1(e) gives an example of our dynamic captioning.

The dynamic captioning is accomplished by exploring a diverse set of technologies, including face detection and recognition, lip motion analysis, visual saliency analysis, etc. The scheme mainly contains three components: script location, script-speech alignment, and voice volume estimation. Script location determines the region in which scripts will be presented. It first performs a script-face matching to establish the speaking face for each piece of scripts (i.e., establish the person from whom the scripts are coming) based on face detection and recognition techniques. It then selects a non-intrusive region around the face via visual saliency analysis in order to avoid the occlusion of important visual content. Script-speech alignment temporally matches each piece of script and the corresponding speech segment, and in this way the scripts can be highlighted word-by-word in synchrony with the speech. Voice volume estimation computes the magnitude of audio signal in a small local window, and visually demonstrates its variation within the scripts.

The main contributions of this paper can be summarized as follows:

(1) We propose a video accessibility enhancement scheme for hearing impaired audience. To the best of our knowledge, this is the first integrated solution to facilitate hearing impaired users in video access.

(2) Our scheme involves the combination of a variety of technologies as well as novel methodologies. For example, the script-face mapping is an important topic per se and our algorithm can be applied to many other applications.

(3) We conduct an in-depth user study to compare different captioning paradigms with real hearing impaired audience. Several conclusions and analysis also shed light on further research in this direction.

The organization of the rest of this paper is as follows. In Section 2, we provide a review on related work. Section 3 introduces the system overview of video accessibility enhancement. In Section 4, we describe the components of the scheme in detail. Experimental results and user study are presented in Section 5. Finally, we conclude the paper in Section 6.

2. RELATED WORK

Hearing impairment refers to conditions in which individuals are fully or partially unable to detect or perceive at least some frequencies of sounds. Efforts on accommodating hearing impaired people in accessing videos can be traced back to 1970s when closed captioning was demonstrated at the First National Conference on Television in Nashville, Tennessee [8]. For television, captions are encoded into Line 21 of the vertical blanking interval in NTSC programming, while teletext (a television information retrieval service developed in the United Kingdom in the early 1970s and closed caption can be transmitted in the teletext signal) is used in captioning transmit and storage in Phase Alternate Line and Sequential Color with Memory. For movie, probably the best-known closed caption in theaters is the Rear Window Captioning System from the National Center for Accessible Media. Other captioning technologies for movie include hand-held displays similar to Personal Digital Assistant, eyeglasses fitted with a prism over one lens and projected bitmap captions. More recently, efforts have also been made to build accessibility features for digital cinemas.

Despite many captioning standards and technologies have been made, the analysis of the impact of captioning on hearing impaired audience is fairly scarce. The earliest study on investigating caption perception of hearing impaired audience shows that adjusting captions to suitable linguistic

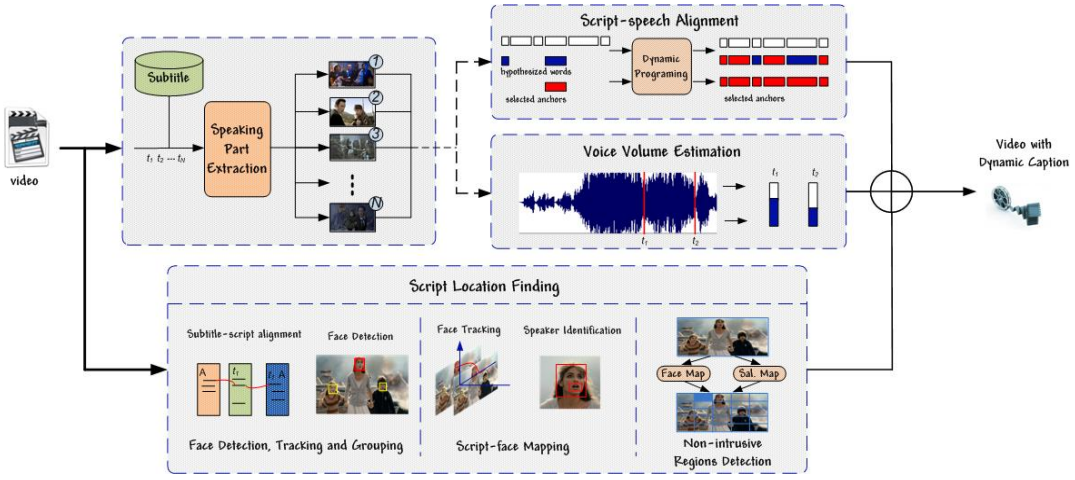


Figure 2: The schematic illustration of the accessibility enhancement.

level and reading rate is able to significantly improve the information gain from captions [10]. Braveman and Hertzog [11] analyzed the language level but not the rate of captioning that affect deaf users' comprehension. Jelinek *et al.* [12] investigated the difference of video caption comprehension between hearing impaired and normal students. Garrison *et al.* [13] studied how working memory affected the language comprehension of deaf students. Gulliver and Ghinea [4, 5] investigated the impact of captions on the perception of video clips for hearing impaired audience. They concluded that much information can be gained from caption, but the information from other sources such as visual content and video text will be significantly reduced, i.e., the caption has no significant effect on the average level of assimilated information across all sources. This indicates that it is not easy for the special audience to track, perceive and learn from the caption efficiently.

Therefore, the existing captioning technology is still far from satisfactory in assisting hearing impaired audience (in Section 1 we have introduced several shortcomings). Recenly, the closed captioning on YouTube is a meaningful exploration towards helping hear impaired users access web videos. There also exists software, such as Captioneer[2], that is able to support manual editing of captions or even add several attractive effects. However, they still can not fully address the aforementioned problems and manual editing is also not an ideal solution due to the high labor cost. In this work we investigate an automatic approach to intelligently present caption. It puts scripts in suitable regions, aligns them with speech and also illustrates the variation of voice volume. The user study with hearing impaired audience has demonstrated the effectiveness of this approach.

3. DYNAMIC CAPTIONING: SYSTEM OVERVIEW

Figure 2 demonstrates the schematic illustration of our video accessibility enhancement process. It mainly contains three components: script location, script-speech alignment and voice volume estimation. Given a video along with its script and subtitle file[3], we first extract speaking parts according to the time information in subtitle. We then map the character faces to the corresponding scripts with face detection and recognition techniques. After that, a non-intrusive region is detected around the face based on visual saliency analysis, in which the scripts are presented.

In parallel, the scripts are aligned with the audio track based on script-speech technology [33], and the starting and ending time of each word are recorded. Based on this information, we synchronously highlight the scripts word-by-word along with the speech so that the hearing impaired audience can better track them. The voice volume estimation component estimates the local power of the audio signal. We visualize it near the scripts to help audience understand the emotion of the corresponding characters.

Our scheme thus generates a set of metadata in addition to the scripts, including the region information of each piece of script, the starting and ending time of each word and the voice volume information. In our work, we use an XML file to record these metadata. With these metadata, we can easily develop an intelligent player to display videos with dynamic caption.

4. DYNAMIC CAPTIONING: TECHNOLOGIES

In this section, we introduce the three components in the proposed scheme in detail.

4.1 Script Location

This sub-section describes our script location algorithm that finds suitable regions for presenting the scripts. As shown in Fig. 2, it comprises three steps: (1) face detection, tracking and grouping; (2) script-face mapping; and (3) non-intrusive region detection.

4.1.1 Face Detection, Tracking and Grouping

In several cases, script file only contain speech content and speaker identity and there is another subtitle file that

[2]http://www.tsstech.org/captioneer.html

[3]In some cases, "subtitle" may assume the audience can hear but cannot understand the language or accent but "caption" aims to describe to the hearing-impaired all significant audio content. In this study, we restrict subtitle as the text that has time information and dialog. There also exists subtitle that contains richer content but it is not easily to acquire.

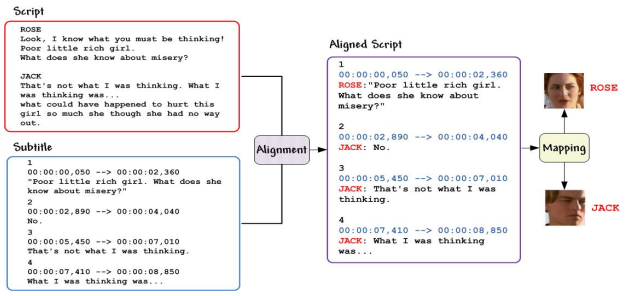


Figure 3: An example of the merge of subtitle and script files. After performing script-face mapping, we can further establish the relationship between script, character identity and faces.

records the time information, as illustrated in Fig. 3. Therefore, we need to merge the speech content, speaker identity and time information from the subtitle and script. Here we utilize a dynamic time warping method [22] to align subtitle and script. Figure 3 demonstrates an example of such alignment. Of course this step can be eliminated if there is only a script file encoding all the information.

Next we implement a face detector to extract faces from the frames in the speaking parts. Here we adopt the face detection algorithm in [17] and several examples of detected faces can be found in Fig. 4(a). As a video may contain thousands or even more detected faces, we group continuously detected faces of a particular character as a face "track" with a robust foreground correspondence tracker [18]. The tracker mainly works as a follows. Given a pair of faces in adjacent frames, the size of overlapped area between the two bounding boxes of faces is estimated. If this value is greater than a given threshold, a matching is declared. This tracking procedure is also able to deal with the cases that faces are not continuously detected due to pose variation or expression change. In this way, the number can be significantly reduced (typically we only need to deal with hundreds of such tracks). As a consequence, face track is adopted as the unit for labeling.

4.1.2 *Script-Face Mapping*

Now we consider the script-face matching problem. The difficulty mainly lies on the following two facts: (1) in many cases there are more than one face within a frame (see the middle image in Fig. 4(a)) and we need to judge who is the speaker; and (2) even when there is only one face in the frame, he/she may not be the speaker and scripts come from another character (the third image in Fig. 4(a) is an example and the girl in the frame is actually not speaking). To deal with these problems, first we adopt lip motion analysis [19] to establish whether the character is speaking when the frame contains only one face based on the fact that speaking is associated with distinctive lip movement.

The lip motion analysis is performed as follows. First we detect a rectangular mouth region within each detected face region using *Haar* feature based cascade mouth detector. Figure 4(b) illustrates several examples of mouth detection. We then compute the mean squared difference of the pixel values within the mouth region between each two continuous frames. To keep translation invariance, the difference is calculated over a search region around the mouth region in the current frame and we then take the minimal difference for decision. Two thresholds are set to establish three statuses, namely "speaking", "non-speaking" and "difficult to judge".

Now we consider the cases that a frame contains more than one face. Our approach is to first label faces with speaker identities and then match them with scripts accordingly (the script file contains the speaker identity information). Note that in cases that the frame only contains one face, we can easily label the face with speaker identity. We then label the face tracks with identities based on such information. For example, if over half of the faces in a track are detected as speaking status and the script shows that merely "EDWARD" is speaking in this period, then we can label this track as "EDWARD" with high confidence. The highly-confident labeled tracks are treated as training exemplars to predict other tracks that are unlabeled due to not containing enough established identities. Each unlabeled face track is simply represented as a set of history image feature vectors. One simple method for identification, as conducted in [22][26], is to directly calculate the feature distance between a testing face track and exemplar face tracks, and then assign testing face track to the nearest neighborhood. Another feasible method is to classify each history image independently via certain classification methods such as sparse representation based classification [20][29][43], and then assign the face track to the class that achieves the highest frequency.

In this work, by regarding the identification of each history image in a testing face track as a task, we formulate the face track identification challenge as a multi-task face recognition problem. This motivates us to apply the multi-task joint sparse representation model [37] to accomplish the task. The key advantage of multi-task learning is that it can efficiently make use of the complementary information embedded in different sub-tasks. We construct the representation of face appearance by a part-based descriptor extracted around local facial features [22]. Here we first use a generative model [39] to locate nine facial key-points in the detected face region, including the left and right corners of two eyes, the two nostrils and the tip of the nose and the left and right corners of the mouth. Figure 4 illustrates the detected key-points of several faces as examples. We then extract the 128-dim Sift descriptor from each key-point and concatenate them to form a 1152-dimensional face descriptor (SiftFD).

The employed face detection, tracking as well as speaker detection are able to offer a number of face tracks where the proposed identity is correct with high probability. For tracks which contain only a single identity, they can be treated as exemplars for labeling other tracks that contain no, or uncertain proposed identity. Each unlabeled face track is, nevertheless, simply represented as a set of history image vectors. For such history image in the track, the identification can be efficiently done via sparse representation classification [20].

Our multi-task joint sparse representation model works as follows. Suppose that we have a set of exemplar face tracks with M subjects. Denote $X = [X_1, \ldots, X_M]$ as the feature matrix in which the track $X_m \in \Re^{d \times p_m}$ is associated with the m-th subject consisting of p_m samples. Here d is the dimensionality of feature and $\sum_{m=1}^{M} p_m = p$ is the total number of samples. Given a testing face as an ensemble of L history images $y^l \in \Re^d$, we consider a supervised L-task linear representation problem as follows:

$$y^l = \sum_{m=1}^{M} X_m w_m^l + \varepsilon^l, l = 1, ..., L \quad (1)$$

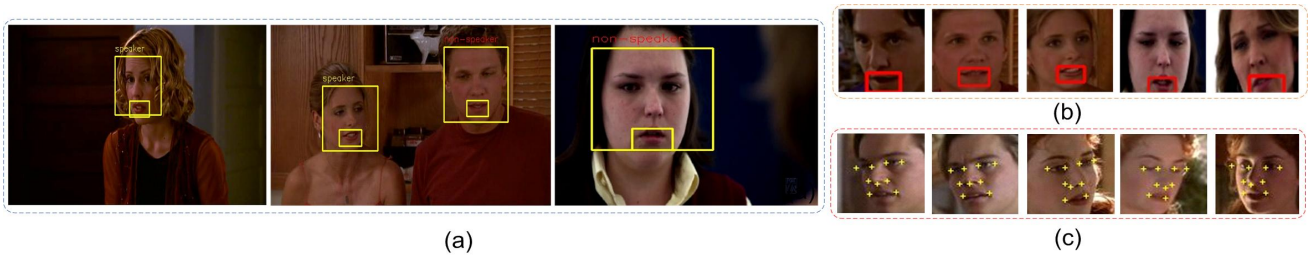


Figure 4: (a) and (b) illustrate the examples of detected faces and mouths, and (c) illustrates the facial feature points of several exemplary frames that are used in multi-task joint sparse face recognition.

where $w_m^l \in \Re^{p_m}$ is a reconstruction coefficient vector associated with the $m-th$ subject, and ε^l is the residual term. Denote $w^l = [(w_1^l)^T, ..., (w_M^l)^T]^T$ as the representation coefficients for probe image feature y^l, and $w_m = [w_m^1, ..., w_m^l]$ as the representation coefficients from the m-th subject across different case images. For simplicity, we denote W as $[w_m^l]_{m \times l}$. Our proposed multi-task joint sparse representation model is formulated as the solution to the following multi-task least square regressions with $\ell_{1,2}$ mixed-norm regularization problem:

$$\min_W F(W) = \frac{1}{2} \sum_{l=1}^{L} \left\| y^l - \sum_{m=1}^{M} X_m w_m^l \right\|_2^2 + \lambda \sum_{m=1}^{M} \|w_m\|_2 \quad (2)$$

Here we use the Accelerated Proximal Gradient (APG) approach [38] to solve the optimization problem in Eqn. (2).

When the optimum $\hat{W} = [\hat{w}_m^l]_{m,l}$ is obtained, a testing image y^l can be approximated as $\hat{y}^l = X_m \hat{w}_m^l$. For classification, the decision is ruled in favor of the class with the lowest total reconstruction error accumulated over all the L tasks:

$$m^* = \arg\min_m \sum_{l=1}^{L} \left\| y^l - X_m \hat{w}_m^l \right\|_2^2 \quad (3)$$

After labeling each face track with speaker identity, we can establish the speaking character even there are more than one face in a frame. Hitherto we have accomplished the mapping between scripts and faces. It is worth mentioning that there also exist scripts that cannot be successfully mapped to faces, and in this work we directly display them on the bottom of frames just like static captioning (off-screen voice is also processed in the same way).

4.1.3 Non-intrusive Region Detection

Up to now, we are able to establish the speaking of each piece of scripts. As previously mentioned, our target is to present the scripts near the speaking face such that hearing impaired audience can easily identify the character from whom the scripts come from. However, we also need to select a region that will not occlude important visual content and especially other faces. Therefore, we perform a visual saliency analysis to select the non-salient regions.

Given an Image I, the contrast of each pixel is an accumulated *Gaussian distance* between it and its neighbors:

$$c_{i,j} = \sum_{q \in \Theta} d(I_{i,j}, q) \quad (4)$$

where $I_{i,j}$ is the pixel position in I and Θ is the neighborhood of $I_{i,j}$. The contrasts $c_{i,j}$ thus form a saliency map [23][25]. Figure 5(b) shows an example of the saliency map

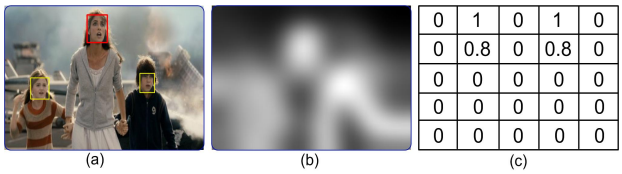


Figure 5: An example of the saliency map and face weighting map for an image from the movie "2012". (a) is the original image; (b) illustrates the saliency map; and (c) shows the weighting map around the speaking face (although there are three faces in the frame, only the face indicated by the red box is speaking).

of the image in Figure 5(a) where the distance is measured in LUV color space. The brighter the pixel in the saliency map, the more important or salient it is.

For the detection of the non-intrusive regions, I is represented by a set of blocks $\mathcal{B} = \{b_i\}_{i=1}^{N_b}$ which are obtained by partitioning image I into $M \times M$ grids ($N_b = M^2$). Each grid corresponds to a block b_i and it gives a candidate region of caption insertion. For each block b_i, a saliency energy s_i ($0 \leq s_i \leq 1$) is computed by averaging all the normalized energies of the pixels within b_i. As previously analyzed, the region should be selected around the speaking face. Therefore, a face weighting map $W = \{w_i\}_{i=1}^{N_b}$ is designed to weight the energy s_i, so that the caption will be restricted around the face. The face weighting map is generated by simply assigning the blocks around the speaker's face block constant weights and all other regions are assigned weight 0. More specifically, the weights of the left and right regions around the face region are set to 1, and the weights of the upper-left, bottom-left, upper-right and bottom right regions are set to 0.8. Figure 5(c) shows an example of the weighting map. Hence, the score for region selection is given by:

$$P(b_i) = w_i \times (1 - s_i) \quad (5)$$

The region with maximal score is finally established for caption insertion. In our work the parameter M is empirically set to 5, but it is also found that an adaptive setting of the parameter will be able to improve performance.

It is worth mentioning that although the non-intrusive region detection approach is effective, it cannot fully guarantee that informative visual content will not be occluded by caption. Thus in dynamic caption we choose to overlay the scripts with parent background such that audience can still recognize the content behind the caption.

4.2 Script-speech Alignment

This section describes our script-speech alignment approach.

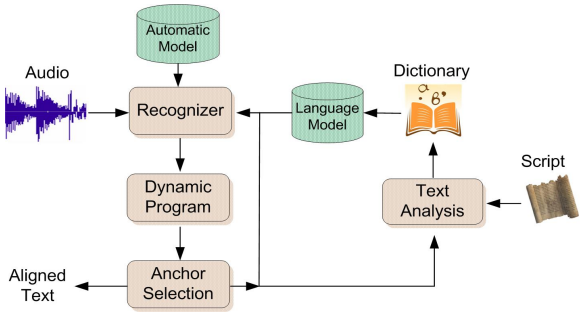


Figure 6: The schematic illustration of the script-speech alignment.

As previously mentioned, based on this component we can synchronously highlight the scripts word-by-word and help impaired audience better track the scripts. Here we adopt a method based on recursive speech recognition with a shrinking dictionary and language model, which is analogous to the approach in [33]. Figure 6 demonstrates a schematic illustration of our scheme. We use 39-dimensional MFCC features. The text analysis module processes the text file and we use the CMU pronouncing dictionary to translate each word into a phonetic sequence. For those words that are out of the dictionary, we use an automatic module introduced in [34] to process them. To reduce the computation cost, we build a simple bigram and trigram word model instead of a complete language model based on N-gram. We then use SPHINX II [35], a speaker-independent speech recognition engine, to recognize the speech based on the previously generated language model and dictionary. When a complete hypothesis text string is produced for the whole audio stream, we employ dynamic programming to find the globally optimum alignment. The detailed process is as follows. We compare the scripts and the recognition results and the matched parts that contain more than N words are regarded as anchors. In our work we empirically set N to 3.

We then iterate the algorithm on each unmatched segment. In each iteration, the language model and dictionary are rebuilt to limit the list of active words and word sequence to those found in the script of this segment. This can speed up the recognition as we only search for those words and their word pairs and triples that are available in the segment. These steps are repeated on the unmatched segments until all the texts have been matched. The iteration also terminates if the recognizer is unable to find any additional words in the audio segment. Our test on 20 video clips (the data are described in Section 5) shows that this approach is able to obtain accuracy, i.e., the ratio of correctly aligned words, of above 90%.

4.3 Voice Volume Analysis

Existing studies reveal that the variation of voice volume conveys important information about human emotion [24, 27]. However, for hearing impaired audience, the volume information is fully lost. Therefore, in our dynamic captioning scheme we symbolize and illustrate the voice volume to help the special audience get more information.

We estimate the sound volume by computing the power of the audio signal in a small local window (the size of the window is set to 30ms in our scheme). After a normalization process, the estimated volume is visualized near the scripts with an "indicator". Figure 1(e) illustrates an example. The volume is indicated by the highlighted part of a strip and the size of the part will vary according to the estimated power.

5. EXPERIMENTS

In this section, we conduct experiments to evaluate the effectiveness and usefulness of the proposed scheme.

5.1 Evaluation of Script-Face Mapping

Our experiments involve 20 clips from three movies, namely *"Titanic"*, *"Twilight"* and *"Up in the Air"*, and one teleplay, namely *"Friends"*. Table 1 presents the information about these clips.

For script-face matching, we have proposed a novel algorithm, namely multi-task joint sparse representation classification. We thus compare it against two existing methods: (1) nearest-neighbor (NN) classifier; and (2) the sparse representation (SR) classifier [20]. For each clip, we use the labeled exemplar faces with high confidence as the training set, and all the detected face tracks are regarded as the testing set. The parameter λ in Eqn. (2) is set to 0.1 throughout the experiment. Figure 7 shows several exemplar training faces and testing face tracks. The accuracies of script-face mapping achieved by our proposed algorithm and two existing methods are given in Table 1. We can see that our proposed algorithm outperforms the other two methods on 15 out of the 20 clips. We can also see that for most clips the recognition accuracy is above 80%. This is important for our scheme, as putting scripts around an incorrect face will be misleading for hearing impaired audience.

5.2 User Study

There are 60 anonymous hearing impaired users participating in the study (21 male and 39 female). These participants come from Huangshan Branch, Anhui Special Education School, China. Their ages vary from 11 to 22. Most of them are pre-lingual deafness which means that they sustained hearing impairment prior to the acquisition of language and occur as a result of a congenital condition or through hearing loss in early infancy. Sign language is their first or preferred language. A small part of participants are post-lingual hearing impaired who occur as a result of disease, trauma or as a side-effect of medicine after the acquisition of language. In our study, two teachers from a deaf-mutes school helped us to communicate with the participants. Before the study, all the participants were required to carefully read the investigation questionnaire and made sure that they understood their roles in the experiment.

We compare the following three paradigms:

(1) No Caption (**NC**), i.e., the hearing impaired participants were shown videos without caption.

(2) Static Caption (**SC**), i.e., the hearing impaired participants were shown videos with static caption (here we adopt the cinematic captioning).

(3) Dynamic Caption (**DC**), i.e., the hearing impaired participants were shown videos with dynamic caption.

We randomly divide all the participants into 3 groups (each group has 20 participants) to avoid the repeated playing of a video which will cause knowledge accumulation[4]. Therefore, each group merely evaluates one of the three paradigms for each video clip.

[4]It is worth noting that we cannot let a user to directly com-

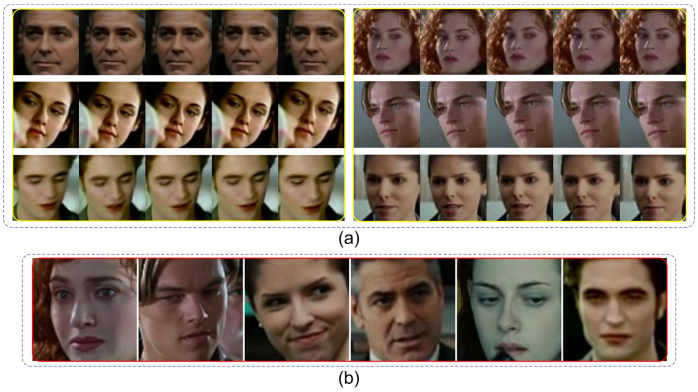


Figure 7: Examples of the selected face tracks and exemplar faces. Five representative images for each track are presented in (a) and the selected exemplar faces with high confidence scores are illustrated in (b).

Table 1: The information about the video clips and the script-face mapping accuracy.

Movie Name	Clips	Frames	Face Tracks	Accuracy (%)		
				NN	*SR*	Proposed Algorithm
"Titanic"	C_1	2,864(1.99min)	22	73.33	72.26	**76.19**
	C_2	7,449(5.17min)	31	86.90	**90.90**	90.24
	C_3	2,868(1.99min)	18	80.89	87.47	**93.75**
	C_4	7,022(4.88min)	22	83.74	**91.51**	88.89
	C_5	9,801(6.80min)	43	**95.00**	**95.00**	87.50
"Twilight"	C_6	4,543(3.15min)	42	88.21	88.21	**89.29**
	C_7	8,193(5.69min)	47	**81.89**	80.60	81.45
	C_8	5,788(4.02min)	51	93.60	93.10	**95.89**
	C_9	6,317(1.95min)	47	75.30	81.90	**86.50**
	C_10	4,745(3.30min)	24	**96.15**	**96.15**	**96.15**
"Up in the Air"	C_11	9,707(6.74min)	22	**100**	**100**	**100**
	C_12	7,955(5.52min)	87	89.01	90.20	**92.52**
	C_13	3,852(2.67min)	31	91.7	**93.55**	92.98
	C_14	6,285(4.36min)	50	91.94	90.45	**92.20**
	C_15	6,533(4.54min)	44	92.75	93.33	**94.40**
"Friends"	C_16	4,549(3.16min)	38	74.26	74.07	**77.78**
	C_17	5,779(4.01min)	26	90.56	92.41	**100.0**
	C_18	3,621(2.51min)	39	78.08	78.61	**80.66**
	C_19	5,036(3.50min)	44	61.52	68.03	**71.90**
	C_20	4,748(3.30min)	31	74.07	86.41	**89.10**

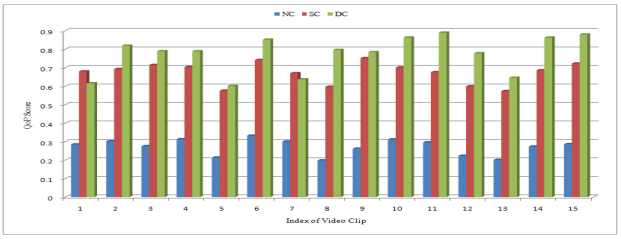


Figure 8: The QoP scores of: (1) no caption; (2) static caption; and (3) dynamic caption. We can clearly see the superiority of dynamic caption.

During the video playing process, participants were informed to stop and answer a number of questions which are related to the content of the movie clips after each showing. To sufficiently investigate the effectiveness of dynamic captioning, we first measure how much advantage our proposed scheme is able to gain on content comprehension and user impression, and then we further evaluate the components. Here content comprehension indicates the extent of understanding from the hearing impaired participants and user impression reflects whether the presentation of such dynamic caption is enjoyable and natural.

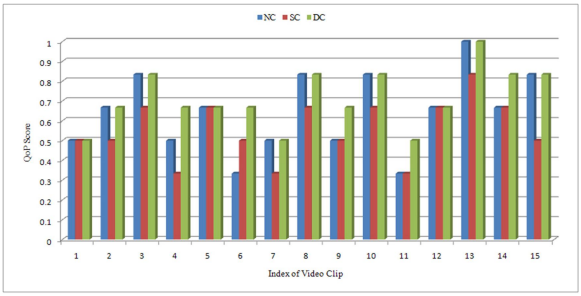
5.2.1 Evaluation of Full Scheme

1. Content Comprehension

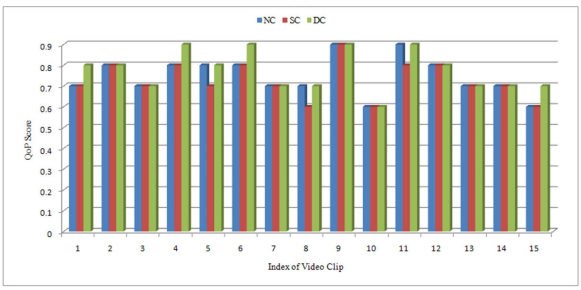
As we know, some questions such as "how many characters are there in this movie clip" have a single definite answer. Thus it is possible to estimate the ratio of correctly answered questions for a pre-defined question set. In our study we have designed 50 questions for each movie clip. These questions are carefully designed to broadly cover the content in the video clip.

pare two paradigms in this study, as the user will get accumulated knowledge if he/she watches the video for more than one time. Thus we divide the participants into groups, and fortunately we have fairly sufficient participants and statistical test (we adopt one-way ANOVA test) can demonstrate the impact of users and the difference of paradigms.

(a)



(b)



(c)

Figure 9: The QoP scores of: (1) no caption; (2) static caption; and (3) dynamic caption. We can again clearly see the superiority of dynamic caption.

The questions can also be categorized according to the information source of their answers. For example, the question "who wore the sports clothes numbered 23?" can only be answered based on the video text information in the video, while "what's the name of hero" can merely be answered based on caption information. Therefore, we can also estimate the ratio of correctly answered questions that are related to different information sources. Thus we categorize the questions as follows:

(1) Caption related: information from the captions only (34 questions in total).

(2) Video Text related: textual information contained in video but not in the caption (6 questions in total).

(3) Visual Content related: visual information contained in movie (10).

We can see that most questions (34 among 50) are related to caption. This is because caption is paramount to understanding the story of video. Hearing impaired participants were asked to answer the questions independently. For performance evaluation, we take the metric of Quality of Perception (QoP), which is defined as the ratio of the correctly answered questions in the full question set.

Table 2: The ANOVA test results on comparing DC and NC. The conclusion is that the difference of the two schemes is significant, and the difference of users is insignificant.

The factor of schemes		The factor of users	
F-statistic	p-value	F-statistic	p-value
86.75	2.47×10^{-11}	0.532	0.818

Table 3: The ANOVA test results on comparing DC and SC. The conclusion is that the difference of the two schemes is significant, and the difference of users is insignificant.

The factor of schemes		The factor of users	
F-statistic	p-value	F-statistic	p-value
32.27	1.93×10^{-11}	0.23	0.971

Figure 8 gives average QoP scores of each video clip (the scores are averaged over participants) with different captioning paradigms. From the figure we can see that both the static (SC) and the dynamic (DC) caption can greatly improve the comprehension level in comparison with the NC paradigm. It is worth noting that this does not contradict with the study in [4][5] which reports that SC can hardly improve the information gain of hearing impaired audience, as in our study most questions are related to caption and thus audience cannot answer these questions without watching the captions. Next we can see that for most clips the DC paradigm outperforms SC. Only for clips C-1 and C-7 the DC paradigm performs slightly worse. This is mainly due to the relatively low script-face mapping accuracies (76.19% for clip C-1 and 81.45% for clip C-7). We also perform a one-way ANOVA test [36], and the results are illustrated in Table 2 and Table 3. From the results we can see that the superiority of DC is statistically significant and the difference among users is statistically insignificant.

We then estimate the QoP scores for different question sets. Figure 9 illustrates the results. We can see that for questions that are related to caption (Figure 9(a)), the performance of NC is very poor as expected, and SC and DC are very close. This indicates that the conversion from static to dynamic caption doesn't add much information. Figure 10(b) shows that the QoP scores of DC are remarkably higher than SC for the questions that are related to video text or visual content. The QoP scores of SC are even much worse than NC for the question related to video text. This indicates that the conventional captioning styles are rather distracting, and the hearing impaired participants need to focus on both the visual content and the caption at the bottom of frames. Our dynamic captioning scheme can be more easily glimpsed as the scripts are presented around the character faces.

Overall, it is clear that captioning is important for understanding the story in videos, but the conventional static captioning approach will degrade the information assimilation of hearing impaired audience from other sources, and our dynamic captioning scheme can help them better perceive the content.

2. User Impression

For user impression, we compare static captioning and dy-

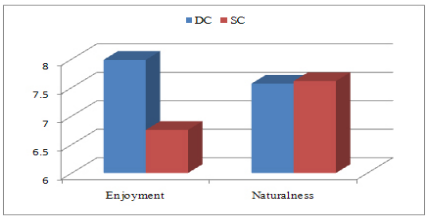


Figure 10: Study results of user impression. For the two criteria, namely *Enjoyment* and *Naturalness*, each user has been asked to assign a score between 1 and 10. Here we have demonstrated the scores averaged over users and video clips.

namic captioning with the following two criteria: *enjoyment* and *naturalness*.

- **Enjoyment**. It measures the extent to which users feel that the video is enjoyable.
- **Naturalness**. It measures whether the users feel the visual appearance of caption is natural.

In this test, we do not need to divide the users into groups, and thus each user was asked to assign a score of 1 to 10 (higher score indicates better experience) to the above two criteria. Figure 10 shows the results that are averaged over video clips and users. We can see that the dynamic caption remarkably outperforms static captioning in terms of enjoyment. However, the naturalness scores of the two captioning schemes are close. Via communicating with the audience, it is found that this is due to the fact that in several cases the regions of script presentation vary abruptly. One possible solution to address this problem is to smooth the variation of the regions for presenting the scripts.

3. Preference between Static and Dynamic Caption

Finally, we ask each user to choose between the static captioning and dynamic captioning that he/she prefers and wishes to use in the future considering all the above factors. The results show that 53 among the 60 users choose dynamic captioning. The remained 7 users choose static mainly because they have already been familiar with static captioning. This clearly demonstrates the usefulness of our scheme.

5.2.2 *Component Evaluation*

Now we further evaluate the components in the dynamic captioning scheme. We compare the following paradigms:

(1) Dynamic captioning (DC), i.e., hearing impaired participants were shown videos with dynamic caption.

(2) DC without volume demonstration (DC-VD), i.e., we remove the voice volume demonstration from the dynamic captioning.

(3) DC without volume demonstration and synchronous highlight (DC-VD-SH), i.e., we remove both voice volume demonstration and script synchronous highlight from the dynamic captioning.

(4) Static captioning (SC), i.e., hearing impaired participants were shown videos with static caption.

Analogous to the previous study, we test the content comprehension of the audience with the above paradigms. We

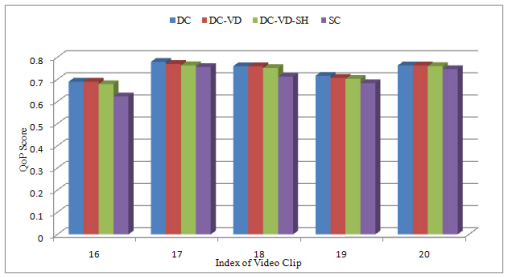


Figure 11: The comparison of QoP scores of: (1) dynamic captioning; (2) dynamic captioning without volume demonstration; (3) dynamic captioning without volume demonstration and script highligh; and (4) static captioning.

conduct this study with the remaining 5 video clips (C-16 to C-20) and for each video we also design 50 questions. We randomly divide the 60 participants into four groups and then implement the question-answering test, and the process is the same with the previously introduced study. Figure 11 illustrates the average QoP scores for different clips. We can see that removing volume demonstration and synchronous highlight will reduce QoP scores, but DC-VD-SH is still able to outperform SC. This demonstrates the effectiveness of each component.

5.3 Discussion

In our experiments, it costs less than 4 minutes to process a video clip on average on a PC with Pentium 4 3.0G CPU and 2G memory. The average duration of the 20 video clips is 3.96 minutes, and this means that the processing time is roughly equivalent to the video duration (of course the processing time also depends on many factors such as the number of characters and the appearing frequency of dialogues). However, it is found that the cost can still be significantly reduced, such as by speeding up the solution process of Eqn. (2) and visual saliency analysis.

We would like to mention that in this work we mainly focus on the technical part of dynamic captioning and care less about user interface, such as the visualization of volume variation (currently we just use a very simple stripe with a highlighted part, see Fig. 1(e)) and the style of script highlight. However, even with a simple interface, our scheme has shown clear advantages through the study of user impression. User interface design is beyond the schope of this paper although it is crucial for real-world application. We will leave it to our future work. Another problem worth mentioning is that inaccurate face-scripts mapping (though only few seen from the Table 1) will not confuse hearing impaired audience in this scheme since these scripts without accurate mapping will be displayed as static captions. This means that to some extent, static captioning can be viewed as the baseline of our scheme.

Finally, we want to emphasize that although the focuses of our scheme are on videos along with scripts, it can be extended to process general videos without scripts. Actually what we need is to employ speech recognition engine to convert speech to scripts, and use speaker clustering [40][41] and identification [42][44] to replace the face grouping and recognition techniques in our current scheme. Of course this task will be much more challenging, but it will be an important topic along this research direction.

6. CONCLUSION

This paper describes a dynamic captioning scheme to enhance the accessibility of videos towards helping hearing impaired audience better enjoy videos. Different from the existing static captioning methods, dynamic captioning put scripts at suitable positions to help hearing impaired audience better recognize the speakers. It also synchronously highlights the scripts by aligning them with the speech signal and illustrates the variation of voice volume to help hearing impaired audience better track and perceive scripts. Comprehensive user study with 60 real hearing impaired participants has demonstrated the effectiveness of our scheme.

As this is the first work to our knowledge to help hearing impaired individuals better access videos, there is a lot of future work along this research direction. We will further improve the script-face mapping component to further boost the mapping accuracy and we will also investigate the extension of the scheme to deal with videos without script, as introduced in Section 5.3. We also plan to conduct a more comprehensive user study on a larger dataset.

7. ACKNOWLEDGMENTS:

This work is supported by NRF/IDM Program of Singapore, under Research Grants NRF2007IDM-IDM002-047 and NRF2008IDM-IDM004-029.

8. REFERENCES

[1] Deaf. http://en.wikipedia.org/wiki/Deaf

[2] J. Nielsen (ed). Advances in human-computer interaction. *vol 5. Intellect Publishers, Bristol, UK.*, 1995.

[3] E.M. Finney and K.R. Dobkins. Visual contrast sensitivity in hearing impaired versus hearing populations: exploring the perceptual consequences of auditory deprivation and experience with a visual language. *Brain Res Cognit Brain Res,* vol.11, no.1, pp.171-183, 2007.

[4] S.R. Gulliver and G. Ghinea. *How level and type of deafness affect user perception of multimedia video clips.* Springer, 2003.

[5] S.R. Gulliver and G. Ghinea. Impact of captions on hearing impaired and hearing perception of multimedia video clips. *ICME,* 2003.

[6] G. Ghinea and J.P. Thomas. QoS impact on user perception and understanding of multimedia video clips. *ACM Multimedia,* pp. 49-54, 1998.

[7] K. Hapeshi and D. Jones. Interactive multimedia for instruction: a cognitive analysis of the role of audition and vision. *Int J Hum Comp Interact,* vol.4, no.1, pp.49-54, 1992.

[8] A brief history of captioned television. *http://www.ncicap.org/caphist.asp.*

[9] S. Cox, M. Lincoln, J. Tryggvason, J. Nakisa, M. Wells, M. Tutt and S. Abbot. TESSA: a system to aid communication with hearing impaired people. *ACM SIGCAPH conference on assistive technologies,* 2002.

[10] J. Boyd and E.A. Vader. Captioned television for the deaf. *Am Ann Hearing impaired,* vol.117, no.1, pp.32-37, 1972.

[11] B.B. Braveman and M. Hertzog. The effects of caption rate and language level on comprehension of a captioned video presentation. *Am Ann Hearing impaired,* vol.125, no.7, pp.943-948, 1980.

[12] L. Jelinek and D.W. Jackson. Television literacy: comprehension of program content using closed captions for the deaf. *J Hearing impaired Stud Hearing impaired Educat,* vol.6, no.1, pp.43-53, 2001.

[13] W. Garrison, G. Long and F. Dowaliby. Working memory capacity and comprehension processes in hearing impaired reader.. *J Hearing impaired Stud Hearing impaired Educat,* vol.2, no.2, pp.78-94, 1997.

[14] Accessibility. *www.wikipedia.org/wiki/Accessibility .*

[15] Web Accessibility Initiative. *http://www.w3.org/WAI/.*

[16] M. Wang. Accessible Image Search. *Proceedings of the ACM Multimedia (MM'09). Beijing, China.*

[17] P. Viola and M. Jones. Rapid object detection using a boosted cascaded of simple features. *CVPR 2001*

[18] T. Yang, Q. Pan, J. Li and S. Li. Real-time multiple objects tracking with occlusion handling in dynamic scenes. *CVPR2 005.*

[19] K. Saenko, K. Liverscu, M. Siracusa, K. Wilson, J. Glass and T. Darrell. Visual speech recognition with loosely synchronized feature streams. *ICCV 2005.*

[20] J. Wright, A. Yang, A. Ganesh, S. Sastry and Y. Ma. Robust face recognition via sparse representation. *IEEE TPAMI. 31(2):210-227. 2009*

[21] M. Bowman, S.K. Debray and L.L. Peterson. Reasoning about naming system. *ACM Transactions on Programming Languages and Systems,* vol.15, no.5, pp.795-825, 1993.

[22] M. Everinghan, J. Siviv and A. Zisserman. Hello! My name is .. Buffy. Automatic naming of characters in TV videos. *BMVC2006*

[23] Y.F. Ma and H.J. Zhang. Contrast-based image attention analysis by using fuzzy growing. *ACM Multimedia,* 2003.

[24] M. Cu, L.T. Chia, H. Yi and D. Rajan. Affective content detection in sitcom using subtitle and audio. *The International Conference on Multi-Media Modeling 2006.*

[25] M. Wang and H. J. Zhang. Video Content Structuring. *Scholarpedia, 4(8):9431.*

[26] M. Wang, X. S. Hua, J. Tang, R. Hong. Beyond Distance Measurement: Constructing Neighborhood Similarity for Video Annotation. *IEEE TMM, 11(3), 2009.*

[27] M. Xu, J.S. Jin, S. Luo and L. Duan. Hierarchical movie affective content analysis based on arousal and valence feature. *ACM Multimedia* 2008.

[28] *http://www.scholarpedia.org/article/Speech emotion analysis.*

[29] M. Wang, X. S. Hua, R. Hong, J. Tang, G. J. Qi, Y. Song. Unified Video Annotation via Multi-Graph Learning. *IEEE TCSVT, 19(5), 2009.*

[30] P. N. Juslin and P. Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code?. *Psychological Bulletin,* vol.129, pp.770-814, 2003.

[31] K. R. Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication,* vol.40, pp.227-256, 2003.

[32] T. J. Hazen. Automatic Alignment and error correction of human generated transcripts for long speech recordings.. *Intl Conf. on Spoken Language Processing* 2006.

[33] P. J. Moreno. A recursive algorithm for the forced alignment of very long audio segments. *International Conference on Spoken Language Processing* 1998.

[34] W. Daelons and V.D. Bosch. TabTalk : Reusability in dataoriented grapheme-to-phoneme conversion. *European Conference on Speech Communication and Technology* 1993.

[35] X. Huang, F. Alleva, H. Hon, M.Y. Hwuang, K.F. Lee and R. Rosenfeld. The SPHINX II Speech Recognition System: An Overview. *Computer Speech and Language,* 1993.

[36] R. A. Fisher. Statistical methods for research workers. Macmillan Pub Co, 1970.

[37] G. Obozinski, B. Taskar and M.I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Journal of Statistics and Computing,* 2009.

[38] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *SIAM J. of Optimization* 2008.

[39] O. Arandjelovic and A. Zisserman Automatic face recognition for film character retrieval in feature-length films. *CVPR'05,* pp.860-867, 2005.

[40] J. Aimera and C. Wooters. A Robust Speaker Clustering Algorithm. IEEE workshop on automatic speech recognition and understanding. 2003.

[41] T. Stadelmann, B. Freisleben. Unfolding speaking clustering potential: a biomimetic approach. ACM MM 2009.

[42] D. A. Reynolds, T. F. Quatieri and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. Digital Signal Processing. 2000(10): 14-91.

[43] J. Tang, S. Yan, R. Hong, G. J. Qi and T. S. Chua. Inferring Semantic Concepts from Community-Contributed Images and Noisy Tags. ACM Multimedia MM'09. Beijing, China.

[44] V. Wan, W. M. Campbell. Support vector machines for speaker verification and ideantification. IEEE proceedings, 2000