VISEME DEFINITIONS COMPARISON FOR VISUAL-ONLY SPEECH RECOGNITION

Luca Cappelletta, Naomi Harte

Department of Electronic and Electrical Engineering
Trinity College Dublin, Ireland
cappelll@tcd.ie

ABSTRACT

Audio-visual speech recognition (AVSR) involves recognising of what a speaker is uttering using both audio and visual cues. While phonemes, the units of speech in the audio domain, are well documented, this is not equally true for the speech units in the visual domain: visemes. In the literature, only a generic viseme definition is recognised. There is no agreement on what visemes practically imply, and if they are just related to mouth position or mouth movement. In this paper a visual-only speech recognition system is presented, trained using either PCA or optical flow visual features. Recognition rate changes depending on which practical viseme definition has been used. Four viseme definitions were tested and results are analyzed in order to establish which is, within the 4 candidates, the best performing viseme definition.

1. INTRODUCTION

Many authors have demonstrated that the incorporation of visual information into speech recognition systems can improve robustness [1]. Papers have described systems performing audio-visual speech recognition on letters [2], digits [3] or words [4, 5] and in a few cases sentences [4]. In the same way, audio-visual speech corpora evolved from isolated digits [6, 7], to isolated words [8], few words sequences [9, 10, 11], to continuous speech [12, 13, 14]. Continuous speech advantages are a full vocabulary, context base utterance and a full coverage of lip positions and movements.

Increasing the vocabulary size, the speech unit for speech recognition has to pass from a word level, to a sub word level. The natural candidate is the *viseme*. A viseme is defined as a visually distinguishable unit, the equivalent in the visual domain of the phoneme in the audio domain [1]. Despite this general definition, it is not clear what a viseme is and how it can be obtained. Moreover, it is not clear yet if the viseme is simply related to mouth position or, in a more complex way, to lip movements.

There is no agreement even on the total viseme number. In this work 4 different viseme practical definitions were tested, and the total viseme number is different in each of them (from 11 to 15, plus a *silence* viseme).

In order to analyze visemes, a big continuous speech audio-visual database is required. The best datasets, in terms of number of speakers and sentences uttered, are AV-TIMIT [14] and IBM ViaVoice [13]. Currently, none of them is publicly available, so a smaller dataset was used in this work: VIDTIMIT [12].

Thus, the aim of this work is to test a speech recognition system, trained using two different visual features (either PCA or optical flow), using several viseme definitions. Since the focus of this work is on the visual part of speech recognition, visual-only cues were tested. No audio cues were used.

This work is structured as follows: firstly, an overview of practical viseme definitions is given, then two feature extraction techniques are presented, and finally a recognition system based on a HMM is presented.

2. VISEME MAPS

This work is focused on continuous speech and the unit of recognition is thus a viseme. As already stated, in literature visemes have different interpretation and there is no agreement on the way to define them. Actually, two practical definitions are plausible:

- Visemes can be thought in terms of *articulatory gestures*, such as lips closing together, jaw movement, teeth exposure, etc.
- Visemes are derived from the grouping of phonemes having the same visual appearance.

The second definition is the most widely used [1, 15, 16, 17], even though no evidence has been provided that it is better than the first definition [15]. Using the second approach, visemes and phonemes are strictly correlated, and visemes can be obtained using a *map* of phonemes to viseme. This map has to be a *many-to-one* map, because many phonemes can not be distinguished using only visual cues. This is the approach used in this work. Within this approach, there are two possible methods to build a map:

Linguistic viseme classes are defined through linguistic knowledge and the *intuition* of which phonemes might appear the same visually.

Data Driven viseme classes are formed performing a phoneme clustering, based on features extracted from the ROIS.

A data driven method has several advantages. First of all, since most viseme recognition systems use statistical models trained on data, it might be beneficial to automatically learn natural classes from data. Secondly, it can account for contextual variation and differences between speakers (but only if a large database is available) [15]. This is particularly important because the linguistic-based method is usually performed with canonical phonemes in mind, while recognition is done on continuous speech.

All the four maps tested in this work has a quite small viseme number (from 11 to 15, plus silence viseme) similar to 14 classes present in the MPEG-4 viseme list [18]. In other maps the viseme number is much higher, e.g. Goldschen map contains 35 visemes [19].

In the first one, Janet & Margaret group 50 phonemes into 11 visemes in the English language [20] for what they

© EURASIP, 2011 - ISSN 2076-1465

describe “as usual viewing conditions”. A map linking phonemes to visemes is shown in Table (1). In this table visemes are labelled using a letter, from /A to /K and a *silence* viseme has been added, labelled using /S. The last column is a suggested phoneme to viseme mapping for the TIMIT phoneme set. Two phonemes are not listed in the table: /hh/ and /hv/. No specific viseme is linked to them because, while the speaker is pronouncing /hh/ or /hv/, the lips are already in the position to produce the following phoneme. Because of this /hh/ and /hv/ have been merged to the following viseme. The table shows the viseme visibility rank and occurrence rate in spoken English [20]. This map is purely linguistic.

The second map analyzed is proposed by Neti *et al* [16]. This map has been created using IBM ViaVoice databaset [13] and using a decision tree, in the same fashion as decision trees are used in order to create triphoneme models. Thus, because of these two aspects, this map can be considered a mix of linguistic and data driven approach. Neti’s map is composed by 43 phonemes and 12 classes (plus a silence class).

Hazen *et al.* [14] use a data driven approach. They perform bottom-up clustering using models created from phonetically labelled visual frames. The map obtained is “roughly” [14] based on this clustering technique. The reason of this apparent inaccuracy is that the clustering results vary a lot depending on the visual feature used. Hazen *et al.* group 42 phonemes into 14 visemes (plus a silence viseme).

Finally, Bozkurt *et al* [17] created a map using the linguistic approach. They define the phonemes clustering as “done in a subjective manner, by comparing the viseme images visually to assess their similarity” [21]. The map is composed by 15 viseme (plus a silence viseme), and 54 phonemes.

It is not simple a task to compare these maps because the number of viseme class and the number of phonemes clustered are not constant within the four maps. Jeffers clusters 50 phonemes in 11 classes, Hazen 42 phonemes in 14 classes, Neti 43 phonemes in 12 classes and Bozkurt 54 phonemes and 15 classes (silence class and phonemes not included). However, it is clear that some similarities are present, particularly between the Jeffers and Hazen maps. All the maps but Bozkurt have 4 vowel classes, but their composition varies a lot within the maps. On the contrary, Bozkurt map has 7 vowel visemes.

On the other hand, more neat is the situation about consonants; in this case all the maps have a specific class for phonemes {/v/, /f/} and for {/ch/, /jh/, /sh/, /zh/}. Moreover, both Jeffers, Neti and Bozkurt have a specific class for {/b/, /m/, /p/} and {/th/, /dh/}, while Hazen splits the first group in two classes and the second is merged with other phonemes. Aside for this, Hazen map is significantly different for the others, while Jeffers and Neti have an impressive class correspondence.

The major difference within the maps is that the phonemes {/pcl/, /tcl/, /kcl/, /bcl/, /dcl/, /gcl/, /epi/} are not considered in the analysis by Jeffer, Neti and Bozkurt, while they are spread into several classes by Hazen.

As an aside, it is possible to analyze the viseme content in a digits task using these different maps. Many visemes are missing, in fact the viseme used in a digit set are 8 (on 11) using Jeffers map, 9 (on 14) using Hazen map, 10 (on 12) using Neti map and 10 (on 15) using Bozkurt map. This further supports the argument that studying a digits task for visual feature is of limited use.

Viseme	Visibility Rank	Occurrence [%]	TIMIT Phonemes
/A	1	3.15	/f/ /v/
/B	2	15.49	/er/ /ow/ /r/ /q/ /w/ /uh/ /uw/ /axr/ /ux/
/C	3	5.88	/b/ /p/ /m/ /em/
/D	4	.70	/aw/
/E	5	2.90	/dh/ /th/
/F	6	1.20	/ch/ /jh/ /sh/ /zh/
/G	7	1.81	/oy/ /ao/
/H	8	4.36	/s/ /z/
/I	9	31.46	/aa/ /ah/ /ay/ /eh/ /ey/ /ih/ /iy/ /y/ /ae/ /ax-h/ /ax/ /ix/
/J	10	21.10	/d/ /l/ /n/ /t/ /el/ /nx/ /en/ /dx/
/K	11	4.84	/g/ /k/ /ng/ /eng/
/S	-	-	/sil/ /pcl/ /tcl/ /kcl/ /bcl/ /dcl/ /gcl/ /h#/ /#h/ /pau/ /epi/

Table 1: Jeffers phonemes to viseme map [20]. The last viseme, /S is used for silence. The table shows the viseme visibility rank and occurrence rate in spoken English. Originally, phonemes {/pcl/, /tcl/, /kcl/, /bcl/, /dcl/, /gcl/, /epi/} were not included in Jeffers & Barley map, so they have been included in the silence class by the authors.

3. FEATURE EXTRACTION

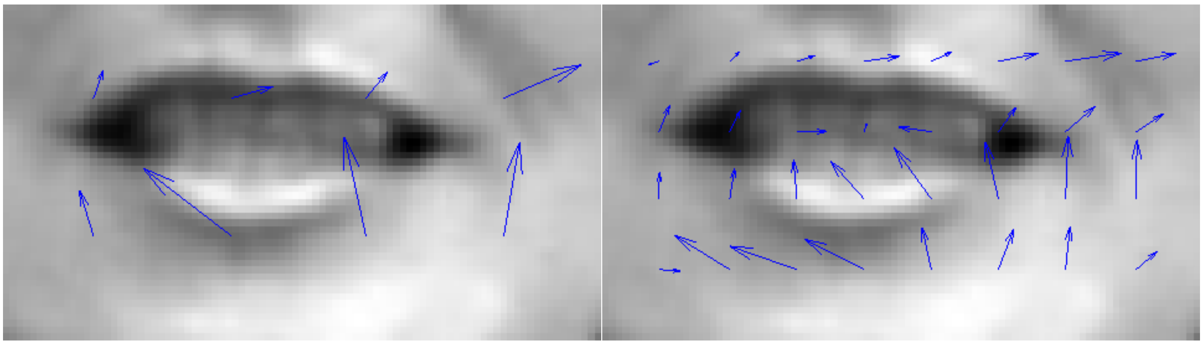
In order to perform feature extraction, the mouth, or ROI (*Region of Interest*), has to be detected. The ROI is found using a technique [22] based on two stages: the speaker’s nostrils are tracked and then, using those positions, the mouth is detected. The first stage succeeds on the 74% of the database sentences, so the remaining 26% has been manually tracked to allow experimentation on the full dataset. The second stage has 100% success rate. Subsequently the ROI is rotated according to the nostrils alignment. At this stage the ROI is a rectangle, but its size might vary in each frame. Thus, ROIs are either stretched or squeezed until they have the same final size. The final size is the mode calculated using all ROIs size.

Having defined the region of interest, a feature extraction stage has to be done, in order to perform the viseme recognition. Two different feature extraction techniques are used: *Principal Component Analysis* (PCA) and *Optical Flow*. Both techniques are appearance-based, rather than shape-based [23].

The first technique used is PCA, also know as *eigenlips* [24]. In this experiment the optimal number of coefficients (the feature vector length) is investigated. A vector too short would lead to a low quality image reconstruction, but a too long one would be difficult to be model with a HMM. Along with PCA features, the use of first and second order derivatives are also investigated. Higher order features are added because it is likely that the lips’ speed and acceleration carry more information than static features. These high order features are defined as:

$$\begin{aligned} D_k[i] &= \mathrm{PCA}_k[i] - \mathrm{PCA}_k[i-1] \\ A_k[i] &= D_k[i] - D_k[i-1] \end{aligned} \qquad (1)$$

where i represents the frame number in the video, and $k \in [1..N]$ represents the kth PCA value. In literature these *dynamic* features are also called Δ and $\Delta\Delta$.



(a) $d_R = 2$ and $d_C = 4$

(b) $d_R = 4$ and $d_C = 8$

Figure 1: Example or two different downsampling.

Optical flow is the distribution of apparent velocities of movement of brightness pattern in a image. It can arise from a relative motion of object or of the viewer. Consequently, optical flow can give important information about the spatial arrangement of the objects viewed and the rate of change of this arrangement [25]. The code used [26] implements the Lucas-Kanade technique [27]. Its output is a two dimensional speed vector for each ROI point. Therefore, a data reduction stage, or *downsampling*, is required. The image is divided in $d_R \times d_C$ blocks, and for each block the median of the horizontal and vertical speed is calculated. In this way $d_R \cdot d_C$ 2D speed vectors are obtained. Figure 1 shows two different downsamplings.

4. EXPERIMENT

4.1 VidTIMIT Dataset

The VidTIMIT dataset [12] is comprised of the video and corresponding audio recordings of 43 people (24 male and 19 female). In each video a single speaker recites a short sentence chosen from the test section of the TIMIT corpus. The selection of sentences in VidTIMIT has a full viseme coverage, no matter what is the viseme definition used. The recording was done in an office environment using a broadcast quality digital video camera at 25 fps. The video of each person is stored as a numbered sequence of JPEG images with a resolution of 512 x 384 pixels. 90% quality setting was used during the creation of the JPEG images. For the results presented in this paper, the database has been divided in a *training* group (295 sentences) and a *test* group (135 sentences). Both groups are balanced in gender and with similar phoneme occurrence rates.

4.2 Hidden Markov Models

A viseme level HMM was trained, using both PCA and optical flow features. A visemic time transcription for VidTIMIT was generated using a forced alignment procedure with monophone HMMs trained on the TIMIT audio database.

The system was implemented using HTK. All visemes were modelled with a left-to-right HMM, except silence which used a fully ergodic model. The number of mixtures per state was gradually increased, with Viterbi recognition performed after each increase to monitor system performance. No language model was used in order to assess raw feature performance. The feature vector rate was increased to 20 ms using interpolation. Both 3 and 4-state HMM were used.

5. RESULTS

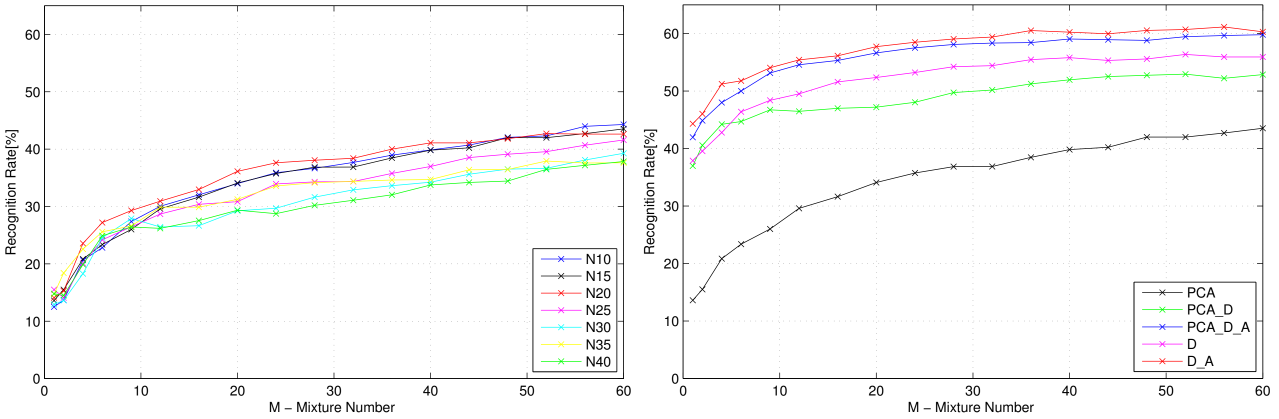
Figure (2) shows the performance of the 3-state HMM using PCA features and the Jeffers map. Results for the 4-state HMM are not shown because no significant improvement from the 3-state was achieved. Figure 2a shows the results of the basic PCA coefficient tests obtained by varying the feature vector length N between 10 and 35. The best performance was achieved with $N = 15$.

Figure 2b shows the performance of 15 PCA coefficients with high order coefficients added. These plots clearly demonstrate the benefit of including dynamic features. Recognition accuracy is at least 40% higher by including both first and second order dynamics (depending on the mixture number). Interestingly this improvement can be achieved by using only dynamic coefficients and leaving out the original PCA features. This support the theory of Bregler *et al.* [24] that “the real information in lipreading lies in the temporal change of lip position, rather then the absolute lip shape”. Across all tests, increasing the gaussian mixtures in each state increases performance. Beyond 35 mixtures the improvement is not significant. $20 \leq M \leq 30$ is a good trade off between recognition rate and system complexity (and computational time).

Figure 3 shows the recognition rate using optical flow features. Once again, the results shown are for 3-state HMMs, using Jeffers phonemes to viseme map. Like for PCA results, the 4-state HMM does not achieve results significantly better than 3-state. The three curves in Figure 3 represent three different downsampling configurations (d_R and d_C values). Unlike the PCA experiment, optical flow performance basically does not change by modifying the experimental setup. Results are virtually the same for all the downsampling configurations. It seems that, no matter what is the downsampling performed, the information carried by this feature is the same. The usage of high blocks number just adds redundancy in the data extracted.

The results of figure 2 and 3 demonstrate that the PCA and optical flow features obtain basically the same visual recognition performance. This would indicate both feature set are capturing broadly similar data on motion in the ROI.

In the second part of the experiments feature set and HMM parameters were fixed, in order to compare the re-



(a) Basic PCA test, 3 States

(b) Higher order features. $N = 15$, 3 States

Figure 2: Viseme recognition rate varying the PCA feature vector length 2a and including higher order dynamics 2b. N10, N15 refer to number of PCA features at 10, 15 etc.. In 2b PCA denotes 15 PCA features only, PCA_D denotes addition of first order dynamics, PCA_D_A denotes inclusions of both first and second order dynamics etc.

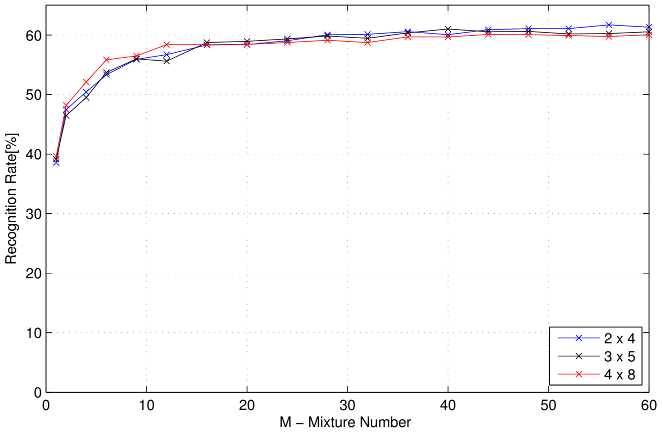


Figure 3: Optical Flow viseme recognition rate varying downsampling values d_R and d_C. Basically, the performance achieved does not vary on downsampling configuration nor state number.

sults from different maps. PCA features were extracted using $N = 15$ and including first and second derivative coefficients only; optical flow features were downsampled using $d_R = 2$ and $d_C = 4$ configuration; only 3-state HMMs have been used.

For this experiment, the optimal number of mixtures for each individual viseme class was tracked.This overcomes issues with different amounts of training data in different classes. Thus HMMs used between 1 to 60 mixtures per state.

As shown in Table 2, Jeffers map obtains the best performance in both PCA and optical flow features. Anyway it is useful to compare Jeffers and Neti maps because, even though many visemes are composed by the same phonemes (5 classes are identical), the results are quite different.

	Viseme recognition rate [%]			
Feature	Jeffers	Hazen	Neti	Bozkurt
PCA	60.1	46.1	46.3	41.8
Optical Flow	57.0	45.6	47.0	44.6

Table 2: Recognition result the on four maps using an optimized mixture number for each viseme.

Jeffers and Neti maps contains respectively 11 and 12 visemes (plus one silence viseme) and in the PCA Jeffers has a 14% improvement over Neti. It is particularly interesting to note that a pure linguistic map achieves a better result than a mixed linguistic and data driven map. It is possible to argue that, even using a large database (IBM ViaVoice database), a linguistic map is still better. However, the maps differ mainly because of the vowel classes. In particular, Jeffers has 2 big and 2 very small vowel classes (in term of number of phonemes contained), while Neti has 4 quite balanced classes (they contains almost the same number of phonemes). Jeffers may have an advantage because misclassification is less probable if classes are big. Moreover, even having a complete misclassification in the 2 small classes, this will have a minor impact on the overall recognition rate. Practically, it is possible to state that Jeffers map has just 2 big vowel classes, because visemes /D and /G (see Table 1) have a very low occurrence and because elements belonging to these classes are usually misclassified as belonging to /B and /I, the other two vowel visemes.

Thus it is possible to see a link between vowel class number and recognition rate. The lower the vowel viseme number, the higher the recognition rate. This basically causes the different performance between Jeffers and Neti maps and it can explain the poor result achieved by Bozkurt map (7 vowel classes).

Jeffers map outperforms even considering the different map guessing rate. Defining it as the reciprocal of total class number, it spans from 8.33% for Jeffers map to 6.25% for Bozkurt map. Jeffers has the highest guessing rate, but the difference from other figures is so small that it can not be considered as the outperform cause.

6. CONCLUSIONS AND FUTURE WORK

This paper has presented a continuous speech recognition system based *purely* on HMM modelling of visemes. A continuous recognition task is significantly more challenging than isolated word recognition task such as digits. In terms of AVSR, it is a more complete test of a systems ability to capture pertinent information from a visual stream, as the complete set of visemes is present in a greater range of contexts.

The importance of dynamic information for visual features is clearly shown as the best performance of 60.1% was achieved using only first and second order PCA derivatives.

Different phonemes to viseme maps have been tested. These maps were created using different approaches (linguistic, data driven and mixed). A pure linguistic map (Jeffers) achieved the best recognition rates, probably because its vowel class configuration.

Work is ongoing to extend this system to comprise other feature sets including different optical flow implementations and *Active Appearance Model* (AMM) features to provide a definitive baseline for visual speech recognition. The emphasis will be given in establish the optimal visual feature set for the capture of dynamics in human mouth movements. Certainly, in order to test Jeffers map effectiveness in a real scenario, the presented tests have to be performed including audio cues as well. We hope to present future results on a larger continuous speech dataset.

Acknowledgement

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant Number 09/RFP/ECE2196.

REFERENCES

[1] G. Potamianos *et al.*, "Recent advances in the autormatic recognition of audio-visual speech," *Proceeding of the IEEE*, vol. 91, no. 9, 2003.

[2] W. C. Yau *et al.*, "Visual Speech Recognition Using Motion Features and Hidden Markov Models," in *CAIP 2007*, S.-V. B. Heidelberg, Ed., 2007.

[3] P. Tsang-Long *et al.*, "Automatic visual feature extraction for Mandarin audio-visual speech recognition," in *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*, 2009, pp. 2936–2940.

[4] S. Alizadeh *et al.*, "Lip feature extraction and reduction for hmm-based visual speech recognition systems," in *Signal Processing, 2008. ICSP 2008. 9th International Conference on*, 2008, pp. 561–564.

[5] Y. Lan *et al.*, "Comparing visual features for lipreading," in *International Conference on Auditory-Visual Speech Processing*, 2009, pp. 102–106.

[6] J. Luettin *et al.*, "Visual speech recognition using active shape models and hidden markov models," 1996.

[7] E. Patterson *et al.*, "CUAVE: a new audio-visual database for multimodal human-computer interface research," in *Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on*, vol. 2, 2002, pp. 2017–2020.

[8] T. Chen, "Audiovisual speech processing," *Signal Processing Magazine, IEEE*, vol. 18, no. 1, pp. 9–21, 2001.

[9] K. Messer *et al.*, "XM2VTSDB: The Extended M2VTS Database," in *Second International Conference on Audio and Video-based Biometric Person Authentication*, 1999.

[10] M. Cooke *et al.*, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[11] N. Fox, B. O'Mullane, and R. Reilly, "The Realistic Multi-Modal VALID database and Visual Speaker Identification Comparison Experiments," in *AVBPA*, New York, 2005.

[12] C. Sanderson, *Biometric Person Recognition: Face, Speech and Fusion*. VDM-Verlag, 2008.

[13] I. Matthews *et al.*, "A comparison of model and transform-based visual features for audio-visual LVCSR," in *Multimedia and Expo, 2001. ICME 2001. IEEE International Conference on*, 2001, pp. 825–828.

[14] T. J. Hazen *et al.*, "A segment-based audio-visual speech recognizer: data collection, development, and initial experiments," in *Proceedings of the 6th international conference on Multimodal interfaces*. State College, PA, USA: ACM, 2004, pp. 235–242.

[15] K. Saenko, "Articulary features for robust visual speech recognition," Master Thesis, Massachussetts Institute of Technology, 2004.

[16] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, S. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Tech. Rep., Oct. 12 2000.

[17] E. Bozkurt, E. Qigdem Eroglu, E. Erzin, T. Erdem, and M. Ozkan, "Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation," in *3DTV Conference, 2007*, 2007, pp. 1–4.

[18] I. S. Pandzic and R. Forchheimer, *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. New York, NY, USA: John Wiley & Sons, Inc., 2003.

[19] A. Goldschen, O. Garcia, and E. Petajan, "Continuous optical automatic speech recognition by lipreading," vol. 1, pp. 572 –577 vol.1, oct-2 nov 1994.

[20] J. Jeffers and M. Barley, *Speechreading (Lipreading)*. Charles C Thomas Pub Ltd, 1971.

[21] T. Ezzat and T. Poggio, "Miketalk: a talking facial display based on morphing visemes," in *Computer Animation 98. Proceedings*, 1998, pp. 96–102.

[22] L. Cappelletta and N. Harte, "Nostril detection for robust mouth tracking," in *Irish Signals and Systems Conference*, Cork, 2010, pp. 239 – 244.

[23] D. Shiell, L. Terry, P. Aleksic, and A. K. Katsaggelos, "Audio-visual and visual-only speech and speaker recognitio: Issue about theory, system design and implementation," in *Visual Speech Recognition: Lip Segmentation and Mapping*, 2008, pp. 1–38.

[24] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," in *Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on*, vol. ii, 1994, pp. II/669–II/672 vol.2.

[25] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artificial Intellicenge*, 1980.

[26] J. Bouguet, "Pyramidal Implementation of the Lucas Kanade Feature Tracker: Description of the algorithm," 2002.

[27] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of Imaging Understanding Workshop*, 1981.