Hindawi Publishing Corporation
EURASIP Journal on Image and Video Processing
Volume 2007, Article ID 45641, 14 pages
doi:10.1155/2007/45641

Review Article

Image and Video for Hearing Impaired People

Alice Caplier,[1] Sébastien Stillittano,[1] Oya Aran,[2] Lale Akarun,[2] Gérard Bailly,[3] Denis Beautemps,[3] Nouredine Aboutabit,[3] and Thomas Burger[4]

[1] *Gipsa-lab, DIS, 46 avenue Félix Viallet, 38031 Grenoble cedex, France*
[2] *Department of Computer Engineering, Bogazici University, Bebek 34342 Istanbul, Turkey*
[3] *Gipsa-lab, DPC, 46 avenue Félix Viallet, 38031 Grenoble cedex, France*
[4] *France Télécoms, R&D-28, Ch. Vieux Chêne, 38240 Meylan, France*

Received 4 December 2007; Accepted 31 December 2007

Recommended by Dimitrios Tzovaras

We present a global overview of image- and video-processing-based methods to help the communication of hearing impaired people. Two directions of communication have to be considered: from a hearing person to a hearing impaired person and vice versa. In this paper, firstly, we describe sign language (SL) and the cued speech (CS) language which are two different languages used by the deaf community. Secondly, we present existing tools which employ SL and CS video processing and recognition for the automatic communication between deaf people and hearing people. Thirdly, we present the existing tools for reverse communication, from hearing people to deaf people that involve SL and CS video synthesis.

1. DEAF PEOPLE LANGUAGES

This section gives a short description of the two possible communication languages used by hard of hearing people.

Sign languages are the primary means of communication of deaf people all over the world. They emerge spontaneously, and evolve naturally within deaf communities. Wherever deaf communities exist, sign languages develop, without necessarily having a connection with the spoken language of the region. Although their history is at least as old as spoken languages, the written evidences showing sign language usage date back to the 16th century, and the earliest record of sign language education dates to the 18th century: in Paris, Abbé de l'Épée founded a school to teach Old French Sign Language and graduated Laurent Clerc who later founded the "Gallaudet College" in U.S. with T. H. Gallaudet. Gallaudet College later became Gallaudet University which is the only liberal art university for deaf people in the world.

The main difference between spoken and sign languages is the way the communicative units are produced and perceived [1]. In spoken languages, the words are produced through the vocal tract and they are perceived as sounds; whereas in sign languages, the signs are produced alone or simultaneously, by use of hand shapes, hand motion, hand location, facial expression, head motion, and body posture, and they are perceived visually. Sign languages have both sequential and parallel nature: signs come one after the other showing a sequential behaviour; however, each sign may contain parallel actions of hands, face, head, or body. Apart from differences in production and perception, sign languages contain phonology, morphology, semantics, syntax, and pragmatics like spoken languages [2]. Figure 1 shows example signs from American Sign Language.

More recently, the Cued Speech language (CS) has been introduced to enrich spoken language by Cornett [3]. The aim of CS is to overcome the problems of lip-reading and to enable deaf people to understand full spoken languages. CS brings the oral language accessible to the hearing impaired, by replacing invisible articulators that participate to the production of sound (vocal cords, tongue, and jaw) by hand gestures, while keeping visible articulators (lips). Basically, it complements the lip-reading by various manual gestures, so that phonemes which have similar lip shapes can be differentiated. Then, considering both lip-shapes and gestures, each phoneme has a specific visual aspect. In CS, information is shared between two modalities: the lip modality (related to lip shape and motion) and the hand modality (related to hand configuration and hand position with respect to the face).

Figures 2 and 3 present the eight different hand shapes which code the consonants and the five hand positions which code the vowels for French CS. CS is different from

Figure 1: American SL examples: from left to right, "sad," "baby," "father," "you."

Figure 2: 8 hand shapes or configurations [FrenchLPCsite].

SL because among others things, CS addresses speech. CS has the same grammar and syntax as current spoken languages (English, French, etc.). For that reason, a deaf person learning English CS learns English at the same time. In contrast, SL of a group of deaf people has no relation to the hearing community of the region apart from cultural similarities.

Figure 4 presents a coding example of the French word "bondir" (to jump).

SL and CS are totally independent and are different languages. Nevertheless, in terms of image and video processing, one can identify some similarities:

(i) both languages involve hand-gesture processing;
(ii) both languages involve face processing;
(iii) both languages are multimodal and need fusion of different modalities.

2. FROM DEAF PEOPLE TO HEARING PEOPLE

CS and SL are two languages based on visual information. Hearing peoples' language is based on speech information (cf., Figure 5). In order to make communication from a deaf person to a hearing person possible, it is necessary to transform the visual information into speech information. Three main steps are necessary for SL or CS automatic translation: SL/CS analysis and recognition, SL/CS to text translation, and speech synthesis. In this review, we concentrate on the first part.

2.1. *Sign language analysis and recognition*

The problem of sign language analysis and recognition (SLR) can be defined as the analysis of all components that form the language and the comprehension of a single sign or a whole sequence of sign language communication. The ultimate aim in SLR is to reach a large-vocabulary sign-to-text translation system which would ease the communication of the hearing and hearing impaired people. The components of a sign, as mentioned above, contain manual signals (MS) such as hand shape, position, and movement, which form the basic components of sign languages, and nonmanual signals (NMS), such as facial expressions, head motion, and

Figure 3: 5 hand positions [FrenchLPCsite]: from left to right, "mouth," "side," "throat," "chin," "cheek bone."

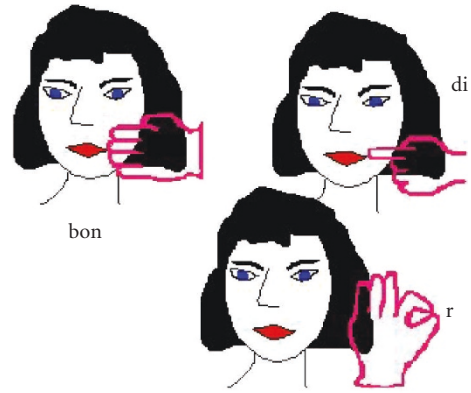


Figure 4: Cued speech coding of the French word "bondir" [FrenchLPCsite].

body posture. A sign-to-text system requires the following components:

(i) hand and body parts (face, shoulders, arms, etc.) detection, segmentation, and tracking;
(ii) analysis of manual signals;
(iii) analysis of nonmanual signals;
(iv) classification of isolated and continuous signs;
(v) natural-language processing to convert classified signs to the text of a spoken language.

Implementing these components provides a system that takes a sign language video and outputs the transcribed text of spoken language sentences. SLR systems can be used in many application areas such as human-computer interaction on personal computers, public interfaces such as kiosks, or translation and dialog systems for human-to-human communication. SLR systems, in connection with sign synthesis, can be used in transferring sign data where the sign is captured at one end and the output of the SLR system can be sent to the other end where it is synthesized and displayed by an avatar. This would require a very low bandwidth when compared to sending the original sign video [4–6]. SLR systems or sign-synthesis systems can also assist sign language education [7–11]. In this review, our focus is on SL analysis and recognition.

2.1.1. *Hand and body parts detection, segmentation, and tracking*

Research on sign language analysis and recognition started with instrumented gloves with several sensors and trackers which provide accurate data for hand position and finger configuration. These systems require users to wear

Table 1: Cues for hand detection and segmentation.

Type of information	Problems	Assumptions/Restrictions
Colour cue	Existence of other skin coloured regions	Long-sleeved clothing
	Contact of two hands	Excluding the face
	Identifying the left and right hands	Only single hand usage
Motion cue	Motion of objects other than the hands	Stationary background
	Fast and highly variable motion of the hand	Hand moves with constant velocity
Shape cue	High degree of freedom of the hand	Restricting the hand shapes

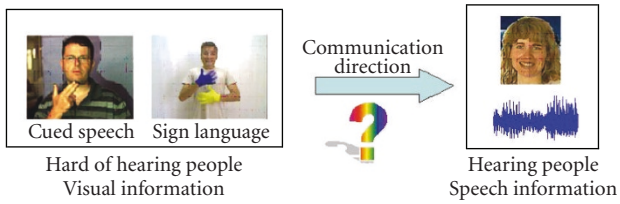


Figure 5: Communication: deaf people to hearing people, from video to speech.

Figure 6: ASL sign "door": markerless hand tracking is challenging since the hands are in front of the face.

cumbersome devices on their hands. Beginning from the mid 90's, improvements in camera and computer hardware have made vision-based hand gesture analysis a possibility [12]. Although these systems provide a natural environment for users, they also introduce several challenges, such as detection and segmentation of hand and finger configuration, or handling occlusion. To overcome some of these challenges, several markers are used in vision-based systems such as different coloured gloves on each hand or coloured markers on each finger. For a brief overview of sign language capturing techniques, interested readers may refer to [13].

Vision-based robust hand detection and segmentation without any markers is still an unsolved problem. Signing takes place in 3D and around the upper body region; thus, the camera field of view must contain the entire region of interest. In a stereo camera setting, both cameras must contain the upper body. An alternative configuration can be using two cameras, one in front, and the other on the right/left side of the signer. More cameras can be used to focus on the face to capture NMS in high resolution.

Colour, motion, and shape information can be used to detect hands in images. However, each source of information has its shortcomings and restrictions (see Table 1). Systems that combine several cues for hand segmentation have fewer restrictions and are more robust to changes in the environment [14–16]. Colour information is used with the strong assumption that hands are the only skin regions in the camera view. Thus, users have to wear long-sleeved clothing to cover other skin regions such as arms [15, 17]. Face detection can be applied to exclude the face from the image sequence, leaving the hands as the only skin regions. However, this approach ignores the situations where the hand is in front of the face: a common and possible situation in sign languages (see Figure 6). When there are two skin regions resulting from the two hands of the signer, the two biggest skin-coloured regions can be selected as the two hands. This approach will fail when the two hands are in contact, forming a single skin-coloured region. Another problem is to decide which of these two regions corresponds to the right and left hands and vice versa. In some studies, it is assumed that users always use or at least start with their left hand on the left and right hand on the right. Starting with this assumption, an appropriate tracking algorithm can be used to track each region. However, when the tracking algorithm fails, the users need to reinitialize the system. Some of these problems can be solved by using motion and shape information. Motion information can be highly informative when the hand is the only moving object in the image sequence [18]. This assumption can be relaxed by combining the motion cue with the colour cue and assuming that the hand is the only moving object among the skin-coloured regions. The main disadvantage of using the shape information alone comes from the fact that the hand is a nonrigid object with a very high degree of freedom. Thus, to achieve high classification accuracy of the hand shape, either the training set must contain all configurations that the hand may have in a sign language video, or the features must be invariant to rotation, translation, scale, and deformation in 3D [19].

Kalman-filter-and particle-filter-based methods are state-of-the-art methods used for tracking the signers' hands. Kalman filters are linear systems with Gaussian noise assumption and the motion of each hand is approximated by a constant velocity or a constant acceleration motion model. Variants of the Kalman filter are proposed to handle nonlinear systems. Two examples are the extended Kalman

Table 2: Feature extraction for Manual signals in vision-based systems.

Hand shape	Hand motion	Hand position w.r.t body
(i) Segmented hand (ii) Binary hand (a) Width, height, area, angle [9, 24] (b) Log polar histograms [19] (c) Image moments [25] (iii) Hand contour (a) Curvature scale space [26] (b) Active contours [15] (iv) 3D hand models	(i) Center-of-mass (CoM) coordinates & velocity hands [9] (ii) Pixel motion trajectory [27] (iii) Discrete definitions of hand motion and relative motion of two hands [19]	(i) Distance to face [9] (ii) Distance to body parts [27] (iii) Discrete body region features [19, 28]

filter and the unscented Kalman filter. Particle filtering, an implementation of which is known as the condensation algorithm (Isard and Blake 1998), is an alternative that works better under nonlinear and non-Gaussian conditions. However, both of these methods need a dynamic model for the hand motion which is not easy to estimate. Apart from these methods, several dynamic programming approaches are also proposed [20].

Hand segmentation and tracking are the most challenging tasks in sign language analysis and recognition. To obtain high recognition rates, an accurate segmentation and tracking is needed. This is possible through the development of methods that are robust to occlusion (hand-hand, hand-face), which frequently occurs in signing.

Besides the hands, several body parts such as the face, shoulders, and arms should be detected [14, 21] to extract the relative position of the hands with respect to the body. This position information is utilized in the analysis of MS. Moreover, the position and motion of the face, facial features, and the whole body is important for the analysis of NMS.

2.1.2. Analysis of manual signals

Manual signals are the basic components that form sign language. These include hand shapes, hand motion, and hand position with respect to body parts (see Table 2). Manual sign language communication can be considered as a subset of gestural communication where the former is highly structured and restricted. Thus, analysis of manual signs is highly connected to hand-gesture analysis [22, 23] but needs customized methods to solve several issues such as analysis of a large-vocabulary system and correlation analysis of signals, and to deal with its structured nature.

For the analysis of hand shapes, a vast majority of the studies in the literature use appearance-based methods. These methods extract features of a hand shape by analyzing a 2D hand image. These features include silhouettes, contours, edges, and image moments, such as Hu or Zernike moments. 2D deformation templates or active contours can be used to find the hand contour. When the position and angles of all the joints in the hand are needed with high precision, 3D hand models should be preferred. However, computational complexity of these methods currently prevents their use in SLR systems.

Some of the studies in SLR literature concentrate only on recognizing static hand shapes. These hand shapes are generally selected from the finger alphabet or from static signs of the language [22, 29, 30]. However, a majority of the signs in many sign languages contain significant amount of hand motion and a recognition system that focuses only on the static aspects of the signs has a limited vocabulary. Hence, for recognizing hand gestures and signs, one must use methods that are successful on modelling the inherent temporal aspect of the data. This temporal aspect can be analyzed as low-level dynamics and high-level dynamics. Modelling low-level dynamics is needed for hand tracking. For this purpose, Kalman-filters- and particle-filter-based methods can be used to estimate the position, velocity, or acceleration of the hand in the next frame given the current frame [9]. High-level dynamics is used to model the global motion of the hand. Sequence-based methods such as hidden Markov models (HMM) [31], Bayesian-network- (BN-) based methods [32], neural-network- (NN-) based methods [27], and temporal templates [33] are used to model the high-level dynamics of the hand. Among these methods, HMMs are used the most extensively and have proven successful in several kinds of SLR systems.

In HMM-based isolated sign recognition approaches, the temporal information of each sign is modelled by a different HMM. For a test sign, the model that gives the highest likelihood is selected as the best model and the test sign is classified as the sign of that model. One of the main challenges is the integration of the two hands and the different features for each hand (shape, motion, position, and orientation) into the HMM model. Different approaches in the literature are summarized in Section 2.1.4.

Grammatical structures in the language are often expressed as systematic variations of the base manual signs. These variations can be in the form of speed, tension, and rate [34, 35]. Most of the SLR systems in the literature ignore these variations. However, special care must be paid to variations especially for continuous signing and in sign-to-text systems.

Table 3: SLR systems using a specialized capture device.

Work	Sign dataset	Capture device	Classification method	Accuracy %*
[42]	5113 CSL signs 750 sentences	Sensored glove & magnetic tracker	Transition movement models (TMM)	91.9
[43]	102 CSL	Sensored glove & magnetic tracker	Boosted HMMs	92.7
[44]	5113 ASL	Sensored glove & magnetic tracker	Fuzzy decision tree, selforganizing Feature Maps, HMM	91.6 SD, 83.7 SI

*SD: signer dependent, SI: signer independent.

Table 4: Vision-based SLR systems.

Who	Sign dataset	Capture restrictions	Hand shape features	Hand motion features	Hand position features	Classification method	Accuracy %
[9]	19 ASL, with NMS	colored gloves	2D app. based	Position, velocity	Distance to face	HMM	99 MS 85 MS + NMS
[15]	21 AusSL, 490 sentences	Dark & static bg., dark long-sleeved clothes	2D geometry based	Movement direction	Geometric features w.r.t face	HMM	99 sign-level 97 sentence-level
[45]	50 ASL, with pronunciation differences	Static background	2D appearance-based features of whole image			HMM with tangent distance	78.5
[24]	439 CSL	coloured gloves	2D appearance-based		Distances of hands to body regions & each other	HMM, Auto-regressive HMM	96.6
[46]	50 ArbSL	coloured gloves	2D binary hand		Hand coords. w.r.t face	HMM	98
[25]	43 BrSL	coloured gloves	Classified hand shape	Movement type	Positions of hands w.r.t body regions & each other	Markov Chains, ICA	97.67
[26]	20 TwSL, single handed	Dark & static bg., dark long-sleeved clothes	2D Curvature scale space	—	—	HMM	98,6 SD

*SD: signer dependent, SI: signer independent.

2.1.3. *Analysis of nonmanual signals*

The nonmanual signals are used in sign language either to strengthen or weaken or sometimes to completely change the meaning of the manual sign. For example, by using the same MS but different NMS, the ASL sign HERE may mean NOT HERE, HERE (affirmative) or IS HERE. The nonmanual signs can also be used by themselves, especially for negation [36, 37]. As opposed to studies that try to improve SLR performance by adding lip reading to the system [38] analysis of NMS is a must for building a complete SLR system: two signs with exactly the same manual component can have completely different meanings. Some limited studies on nonmanual signs attempt to recognize only the NMS without MS. In [39, 40], head movements and in [41], facial expressions in ASL are analyzed. In SLR literature, there are only a few studies that integrate manual and nonmanual signs [9].

2.1.4. *Classification of isolated & continuous signs*

Initial studies on vision-based SLR focused on limited vocabulary systems. These systems can be classified into two groups: those that use vision-based capture methods and those that use device-based capture methods. For a list of SLR systems in the literature, users may refer to the excellent survey in [36]. In Tables 3 and 4, we list several selected SLR systems proposed in the past years that are not covered in [36]. The tables show that almost all of the systems use HMMs or HMM variants for classification. Although the recognition rates of device or vision-based systems are comparable, the shortcoming of vision-based

systems is that they make a lot of assumptions and introduce several restrictions on the capturing environment.

As the vocabulary size increases, computational complexity and scalability problems arise. One of the solutions to this problem is to identify phonemes/subunits of the signs like the phonemes of speech. The advantage of identifying phonemes is to decrease the number of units that should be trained. The number of subunits is expected to be much lower than the number of signs. Then, there will be a smaller group of subunits that can be used to form all the words in the vocabulary. However, the phonemes of sign language are not clearly defined. Some studies use the number of different hand shapes, motion types, orientation, or body location as the phonemes [47, 48]. Others try to automatically define the phonemes by using clustering techniques [49] (see Table 3).

Recognizing unconstrained continuous sign sentences is another challenging problem in SLR. During continuous signing, signs can be affected by the preceding or succeeding signs. This effect is similar to the coarticulation in speech. Additional movements or shapes may occur during transition between signs. These movements are called movement epenthesis [50]. These effects complicate the explicit or implicit segmentation of the signs during continuous signing. To solve this problem, the movements during transitions can be modelled explicitly and used as a transition model between the sign models [15, 42, 48].

2.1.5. Discussion

Isolated SLR achieved much attention in the past decade and systems were proposed that have high accuracies in the reported databases of a wide range of sign languages from all over the world. However these datasets contain different range and number of signs that may be recorded with strong restrictions (slow speed, nonnative signers, unnatural signing, etc.). There are no benchmark sign datasets that researchers can test and compare between their systems. Although there are some publicly available datasets [9, 25, 51, 52], these datasets have not yet become benchmark datasets of SLR researchers.

Current challenges of SLR can be summarized as continuous signing, large vocabulary recognition, analysis and integration of nonmanual signals, and grammatical processes in manual signing. Although these aspects are mentioned by several researchers of the field [36, 53], the amount of research in these areas is still limited. Significant progress can be made by close interaction of SLR researchers with SL linguists.

2.2. Cued Speech analysis and recognition

The problem of Cued Speech recognition involves three main steps: manual-gesture recognition (hand configuration and hand position with respect to the face), lip reading, and hand and lip information fusion in relation with higher-level models to obtain a complete lexical phonetic chain.

Though the number of different hand configurations is less important for CS than for SL and though only a single hand is involved, the problem of hand configuration recognition is a problem which is quite similar to the problem of manual gesture recognition of SL (see Section 2.1.2). As a consequence, we are focusing on hand-position recognition, lip reading, and hand and lip data flow fusion.

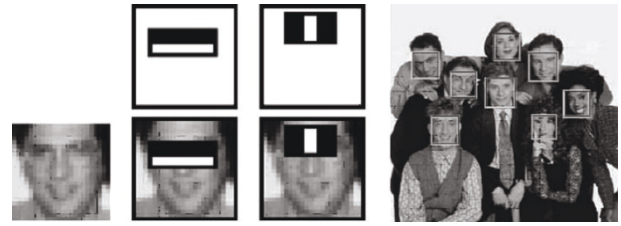


Figure 7: Viola and Jones face detector (from [54]).

2.2.1. Hand-position recognition and face processing

Once the hand configuration (for consonant coding) has been recognized, the second information carried by the coding hand is the position which is pointed by the hand with respect to the face (for vowel coding). In French CS, five positions are considered ("mouth," "side," "throat," "chin," and "cheekbone"). In order to detect the corresponding position, the coder's face and some of his/her facial features have to be detected.

Face localization and facial feature extraction have been extensively studied. Some specific conferences on that topic such as the *IEEE International Conference on Automatic Face and Gesture Recognition* have been created 15 years ago. As stated in [27], the general definition of face detection is the following: "given an arbitrary image, the goal of face detection is to determine whether or not there are any faces in the image and, if present, to return the image location and extent of each face."

The most popular face detector is those developed by Viola and Jones [54] because of its efficiency and because of the accessibility of a downloadable version of the corresponding code [MPT]. This face detector involves an efficient and fast classifier based on Adaboost learning algorithm in order to choose the most discriminating visual features among an important set of potential features. Visual features are based on Haar basis functions. As a result, a bounding box around the detected face is given (see Figure 7).

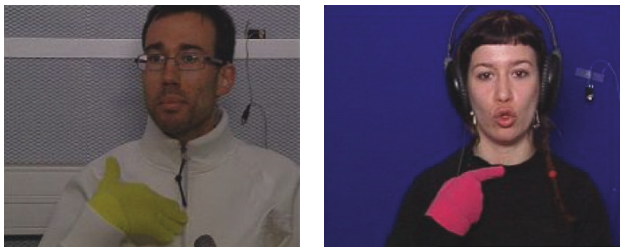
The reader could refer to the two following survey papers about face detection [55–57]. More recently, some new algorithms have been proposed [58–62].

In the context of CS framework, the task of face detection is a little bit simpler since it is assumed that the processed image contains a face and more precisely a frontal-view face. As a consequence, it is not necessary for the used face detector to be robust to different face orientations or to face occlusions.

Most of the face detectors give as output not only the face localization but also the positions of eyes, mouth, and nose. This information is welcome for detecting the position pointing by the coding hand. From morphological and geometrical considerations, it is possible to define 5

Table 5: SLR systems with subunit-based approaches.

Who	Dataset	Phonemes/ Subunits	Capture method	Subunit determination method	Classification method	Accuracy %
[47]	5100 CSL	2439 etymons	Device-based	Manual	HMM	96 sign-based 93 etymon-based
[49]	5113 CSL	238 subunits	Device-based	Automatic, temporal clustering	HMM	90.5
[48]	22 ASL signs 499 sentences	Based on Movement-Hold model	Device-based	Manual	Parallel HMM	95.5 sign-level 87.9 sentence-level



(a) (b)

Figure 8: Different coding for the "throat" hand position.

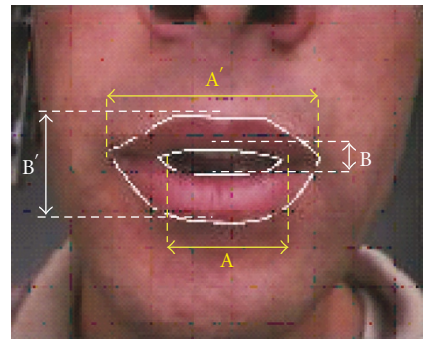


Figure 9: Lip reading parameters.

pointed areas, with respect to these features. However the main difficulty is that for each theoretical position of the CS language, the corresponding pointed area depends on the coder habits. Figure 8 presents an illustration of the differences between two coders for the "throat" position. On the contrary, the pointing area for "mouth" hand position is easier to define once the mouth has been detected.

2.2.2. *Lip reading*

The main difference between SL and CS is that CS message is partly based on lip reading; it is as difficult to read the lip without any CS hand gesture than to understand the hand gestures without any vision of the mouth.

The oral message is mainly encoded in the shape and the motion of the lips. So it is necessary to first extract the contours of the lips and then to characterize the lip shape with an efficient set of parameters. It has been proved that vowels could be characterized with the four parameters presented on Figure 9, the interlabial surface and the mouth's opening surface. These parameters are also used for consonants recognition, but additional information such as teeth or tongue appearing is used. Front views of the lips are phonetically characterized with lip width, lip aperture, and lip area.

The problem of lip reading has historically been studied in the context of audio-visual speech recognition. Indeed, human speech perception is bimodal: we combine audio and visual information. The reader can refer to [63] for a complete overview about audio-visual automatic speech recognition.

In CS, lip reading is associated with CS codes of the hand in a nonsynchronous manner. Indeed, in the case of vowels, the hand often attains the target position before the corresponding one at the lips (see [64], for an extensive study on CS Speech production). As a consequence, the correct identification of the vowel necessitates the processing of the hand flow and the lip flow at two different instants. Thus automatic CS recognition systems have to take into account this delay, and that is why the lip-reading process has to be considered jointly with the hand [65].

The first step is lip contours extraction. Many researches have been carried out to accurately obtain outer lip contour. The most popular approaches rely on the following.

(i) Snakes [66]—because of their ability to take smoothing and elasticity constraints into account [67, 68].

(ii) Active shape models and appearance shape models. Reference [69] presents statistical active model for both shape (AMS) and appearance (AAM). Shape and grey-level appearance of an object are learned from a training set of annotated images. Then, a principal component analysis (PCA) is performed to obtain the main modes of variation. Models are iteratively matched to reduce the difference between the model and the real contour by using a cost function.

(a) (b) (c) (d) (e)

Figure 10: Lip segmentation results.

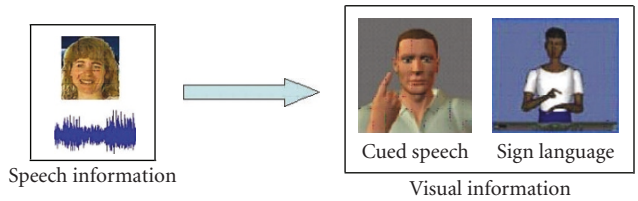


Figure 11: Communication: hearing people to deaf people, from speech to video.

In [70], a parametric model associated with a "jumping snake" for the initialization phase is proposed. For lip reading applications, the accuracy of the lip segmentation is of great importance since labial parameters are then extracted from the segmentation for the purpose of phonems recognition. The model proposed in [70] is the one which is the most adapted to such constraint of accuracy. Figure 10 presents some segmentation results with Eveno's model for the external lip contours.

Relatively few studies deal with the problem of inner lip segmentation. The main reason is that artifact-free inner contour extraction from frontal views of the lips is much more difficult than outer contour extraction. Indeed, one can encounter very distinct mouth shapes and nonlinear appearance variations during a conversation. Especially, inside the mouth, there are different areas such as the gums and the tounge, which have similar color, texture, or luminance than the lips. We can see very bright zones (teeth) as well as very dark zones (the oral cavity). Every area could continuously appear and disappear when people are talking. Among the few existing approaches for inner lip contour extraction, lip shape is represented by a parametric deformable model composed of a set of curves. In [71], Zhang uses deformable templates for outer and inner lips segmentation. The chosen templates are three or four parabolas, depending on whether the mouth is closed or open. The first step is the estimation of candidates for the parabolas by analyzing luminance information. Next, the right model is chosen according to the number of candidates. Finally, luminance and color information is used to match the template. This method gives results which are not accurate enough for lip reading applications, due to the simplicity and the assumed symmetry of the model. In [72], Beaumesnil et al. use internal and external active contours for lip segmentation. In [73], an AMS is built, and in [74], an AMS and an AAM are built to achieve inner and outer lips detection. The success of these models is that the segmentation gives realistic results, but the training data have to contain many instances of possible mouth shapes.

On the contrary, the parametric model described in [75] and made of four cubics is the most accurate lip model for the purpose of lip-reading. According to some optimal information of luminance and chrominance gradient, the parameters of the model are estimated. Figure 10 presents some internal lip contour extraction with Stillittano's algorithm.

2.2.3. Hand and lip merging models

The CS recognition needs to merge both manual and lip flows. The classical models of audio-visual integration and merging in speech (see [63, 76]) have to be considered in the adaptation to the CS. The direct identification model (DI) is based on the classification of a vector of components. This model is not appropriate to mix qualitative and continuous components, as it is particularly the case in CS with hand position, hand shapes, and lip parameters. Even if a transformation of the CS hand in quantitative components is possible, scaling problems arise when components of different nature are combined to form a vector. In the RMD model (recoding in the dominant modality), the auditory modality is the dominant one in speech. The visual modality predicts the spectral structure of the dominant modality. In CS, hand and lip flows are complementary, and thus none of these two modalities can be prioritized. Finally, the separated identification model (the SI fusion of decisions model) seems to be more convenient. In this model, a decision can be made from each flow, independently from the other one. The recognition is the result of the combination of the two decisions. This model has been adapted to the case of CS, for vowel recognition [65, 77]. In the process, the lip parameters at the instant of vowel lip target were used in a Gaussian classification for each of the five CS hand positions. This resulted in a set of five vowels, each one associated to a specific hand position. On the other hand, the hand position is obtained as the result of a classification at the instant of CS hand target [78]. The merging result is obtained as the combination of the decision on the hand position with the set of the five vowels for the (CS hand target instant, lip target instant) couple of instants. As the first result on vowel recognition, an accuracy less than 80% is obtained in the case the two nearest CS hand target and lip target instants are considered under the condition that the CS hand target is ahead of the lips.

3. FROM HEARING PEOPLE TO DEAF PEOPLE

For communication from a hearing person to a deaf person, it is necessary to transform the speech signal into a visual signal; it is necessary to synthesize an avatar producing the associated manual gestures of SL or CS (see Figure 11).

The main step for this communication is the synthesis of the gestural languages.

3.1. General scheme

A synthesis system consists mainly in three components: (1) a trajectory formation system that computes a gestural score given linguistic or paralinguistic information to be transmitted, (2) a shape model that computes the 2D or 3D

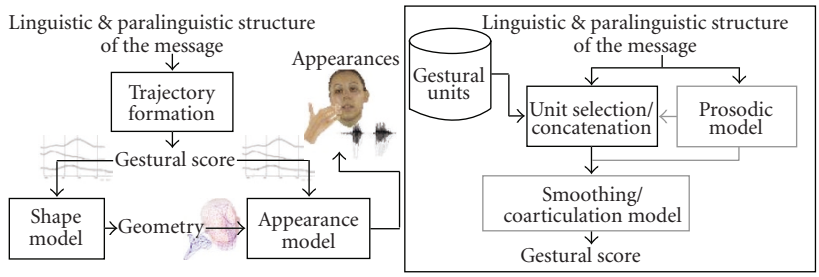


Figure 12: Synopsis of a synthesis system; left: main modules; right: detailed components of the trajectory formation system.

geometry of organs to be moved given the gestural score, and (3) an appearance model that computes pixels to be displayed given this varying geometry.

3.2. Trajectory formation systems

The trajectory formation system is responsible for computing gestural parameters from linguistic and paralinguistic information. This symbolic information consists essentially of basic elements (visemes or kinemes for CS or SL) organized in complex phonological structures that organize the basic left-to-right sequence of these elements in meaningful units such as syllables, words, phrases, sentences, or discourse. These units are typically enriched by unit-specific linguistic information such as accents for the words, dependency relations between phrases, modality for the sentence, and unit-specific paralinguistic information such as emphasis for syllables (narrow focus) or words (broad focus), emotional content for discourse units. The literature often distinguishes between segmental and suprasegmental (often terms as prosodic) units to distinguish between the different scopes. The way suprasegmental information influences segmental gestures is under debate: quantitative models range from simple superpositional models to more complex data-driven functional models. Proceedings of series of conferences on speech prosody (Aix-en-Provence 2001, Nara, Japan, 2004 and Campinas, Brazil, forthcoming in 2008) may provide more detailed insight on these issues.

3.2.1. From gestural units to gestural scores

A trajectory formation system consists thus in three main components (see Figure 12 right): a unit selection/concatenation module, a prosodic model, and a smoothing/coarticulation model. The unit selection/concatenation module selects gestural units according to segmental content. These gestural units can be subsegmental (e.g., gestural strokes) or instances of larger units (triphones or dikeys for CS or complete gestures for SL). Multirepresented units can then be selected using the input suprasegmental information or gestural specification computed by a prosodic model (e.g., computing gesture durations or body and head movements according to wording and phrasing). Note that recently, speech synthesis systems without a prosodic model but using a very large number of gestural units of different sizes have been proposed [79].

Gestural trajectories stored in the segment dictionary are basically parameterized in three different ways: (a) due to memory limitations, earliest systems characterized trajectories by key points (position and velocities at tongue targets in [80], hand targets in [81]), and an organ-specific coarticulation model was used to take into account contextual variations and interpolate between targets; (b) the availability of large video and motion capture resources now give rise to pure concatenative synthesis approaches (e.g., lip motion in [82], or cued speech gestures in [83]) where complete gestural trajectories are warped and smoothed after raw concatenation; (c) more sophisticated statistical models have been recently proposed (such as hidden Markov models in [84]) that capture gestural kinematics and generate directly gestural scores without further processing.

3.2.2. Segmentation and intergestural coordination

Gestural languages encode linguistic and paralinguistic elements and structures via coordinated gestures of several organs: torso, head, arm, and hand gestures, as well as orofacial movements. The trajectory formation system has thus to ensure that these gestures are properly coordinated in order to produce necessary geometric patterns (e.g., signs in SL or mouth shapes and hand/face contacts in CS) at a comfortable cadence for the interlocutor. Reference [64] have thus shown that hand gestures are typically well ahead of lip movements encoding the same syllable.

The unit dictionary is fed by the segmentation of complete discourse units into gestural units, and the choice of gestural landmarks that cue these boundaries is crucial. When considering large gestural units, rest or hold positions can be used as landmarks and intergestural phasing is intrinsically captured within each gestural unit. When smaller units are considered, gestural landmarks (such as targets or maximum velocities) chunking the movements of each organ are not necessarily synchronized and a specific intergestural coordination module should be proposed to generate the appropriate phasing between the different landmarks. Reference [85] have recently proposed a joint estimation of such organ-specific coordination modules and HMM-based trajectory formation systems.

3.3. Shape and appearance models

For nearly 30 years, the conventional approach to synthesize a face has been to model it as a 3D object. In these

Table 6: CS synthesis systems.

Who	Input	Trajectory formation	Shape model
[97]	Speech recognition	Key frames, rule based	2D
[64]	Text	Context-sensitive key frames, rule based	2D
[83]	Text	Concatenation of multimodal gestural units	3D

model-based approaches, control parameters are identified, whick deform the 3D structure using geometric, articulatory, or muscular models. The 3D structure is then rendered using texture patching and blending. Nowadays, such comprehensive approaches are challenged by *image-based* systems where segments of videos of a speaker are retrieved and minimally processed before concatenation. The difference is actually subtle since both approaches underlie at some level the computation of 2D or 3D geometry of organs.

There are mainly three types of image-based systems: (a) systems that select appropriate subimages of a large database and patch selected regions of the face on a background image (see [86, 87]), (b) systems that consider facial or head movements as displacements of pixels (see [88, 89]), and (c) systems that also compute the movement and change of the appearance of each pixel according to articulatory movements [90].

3.4. Text to-gesture systems

Numerous audiovisual text-to-speech systems are still post-synchronizing an animated face with a text-to-speech system that has been developed with separate multimodal resources, often with different speakers [91, 92]. Only recently, systems based on synchronous multimodal data have been developed [93, 94]. When substituting sound with gestures, the difficulty of gathering reliable multimodal data is even more challenging. Contrary to facial gestures where a linear decomposition of shape and appearance can lead to very convincing animations [95], arm and hand movements both in CS and SL require more sophisticated nonlinear models of shape, appearance, and motion [96]. Moreover, acquiring motion capture data or performing video segmentation and resynthesis of articulated organs occluding each other is quite challenging.

The only text-to-CS systems developed so far are presented in Table 6.

3.5. Alternative specifications. Speech-to-gesture systems

The trajectory formation systems sketched above compute movements from symbolic information. This symbolic information can first be enriched with timing information such as delivered as a by-product of unimodal or multimodal speech or gesture recognition. The task of the trajectory formation system is then simplified since prosodic information is part of the system's input. This digital information on the desired gestures can of course include some more precise information on the desired movement such as placement or even more precise geometric specification.

Synthesis systems are also often used in augmented reality systems that superpose gestures of virtual hands or the animation of a virtual signer to videos of a normal hearing interlocutor so that deaf viewers can understand mediated face-to-face conversation. In this case, gestures should be estimated from speech signals and/or video inputs. Reference [98], for example, patented an SL synthesis system that consists in superimposing series of virtual signs, on the original image of the subject to create a synthesized image in which the subject appears to be signing. Duchnowski et al. [81] superimpose an animated virtual hand to the original video of the interlocutor of a speech cuer. The gestures of the virtual hand are computed combining a speech recognizer and a CS synthesizer. The estimation of the phonetic content of the message is however difficult and depends on how much linguistic information—via language models—can be injected into the decoding process. More direct speech-to-gesture mapping has been proposed using conversion techniques. Earliest attempts by [99] demonstrate that a large part of intergestural correlations present in goal-directed movements can be captured by simple linear models. More recently, [100] proposed to use Gaussian mixture models to estimate articulatory motion of orofacial organs from acoustics. The coupling of such mapping tools with state models such as HMMs [101] is worth considering in the future.

3.6. Evaluation

Up to now, there has been no large-scale comparative evaluation campaigns similar to those recently set up for speech synthesis [102, 103]. The multiple dimensions of subjective evaluation (intelligibility, comprehension, naturalness agreement, etc.) together with additional parameters affecting performance (learnability, fatigue, cognitive load) are very difficult to sort out. Despite the fact that Ezzat's talking head Mary [95] passed a first Turing test successfully (do you face a natural or synthetic video?), the visual benefit in segmental intelligibility provided by the data-driven talking face was worse than those provided by the original videos.

A more systematic evaluation was performed at ATT [104] on 190 subjects to show that subjects can compensate for lower intelligibility by increasing cognitive effort and allocate more mental resources to the multimodal decoding.

Recently, Gibert et al. evaluated the segmental and suprasegmental intelligibility of their text-to-CS system with eight deaf cuers [105]. The benefit of cued speech regarding decoding performance of the subjects in a modified Diagnostic Rhyme Test [106] involving pairs of labial sosies in valid French words is impressive. The comprehension of

entire paragraphs is, however, still deceiving; the prosodic cues helping the interlocutor to chunk the stream of gestural patterns in meaningful units are essential to comprehension. Reference [107] has notably shown that early signers are more intelligible than interpreters or late signers and that part of this difference is due to the larger use of body movements.

4. CONCLUSION

Over the last decade, active researches have produced novel algorithms first for improving the communication with deaf people and second to make new technologies accessible to deaf people (TELMA phone terminal, e.g.). These researches are strongly related with the development of new dedicated systems for human computer interaction.

Image processing is used in order to automatically analyze the specific communication languages of deaf people. While the automatic translation of Cued Speech language is now feasible, sign language analysis is still an open issue because of the huge number of different signs and because of the dynamic and 3D aspects of this language, which makes an automatic analysis very difficult.

REFERENCES

[1] S. K. Liddell, *Grammar, Gesture, and Meaning in American Sign Language*, Cambridge University Press, Cambridge, UK, 2003.

[2] W. C. Stokoe Jr., "Sign language structure: an outline of the visual communication systems of the american deaf," Studies in Linguistics: Occasional papers 8, 1960.

[3] R. O. Cornett, "Cued speech," *American Annals of the Deaf*, vol. 112, pp. 3–13, 1967.

[4] R. A. Foulds, "Biomechanical and perceptual constraints on the bandwidth requirements of sign language," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 12, no. 1, pp. 65–72, 2004.

[5] M. D. Manoranjan and J. A. Robinson, "Practical low-cost visual communication using binary images for deaf sign language," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 1, pp. 81–88, 2000.

[6] G. Sperling, "Video transmission of american sign language and finger spelling: present and projected bandwidth requirements," *IEEE Transactions on Communications*, vol. 29, no. 12, pp. 1993–2002, 1981.

[7] Y.-H. Chiu, C.-H. Wu, H.-Y. Su, and C.-J. Cheng, "Joint optimization of word alignment and epenthesis generation for Chinese to Taiwanese sign synthesis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 28–39, 2007.

[8] K. Karpouzis, G. Caridakis, S.-E. Fotinea, and E. Efthimiou, "Educational resources and implementation of a Greek sign language synthesis architecture," *Computers and Education*, vol. 49, no. 1, pp. 54–74, 2007.

[9] O. Aran, I. Ari, A. Benoit, et al., "SignTutor: an interactive sign language tutoring tool," in *Proceedings of the SIMILAR NoE Summer Workshop on Multimodal Interfaces (eNTERFACE '06)*, Dubrovnik, Croatia, July-August 2006.

[10] J. Ohene-Djan and S. Naqvi, "An adaptive WWW-based system to teach British sign language," in *Proceedings of the 5th IEEE International Conference on Advanced Learning Technologies (ICALT '05)*, pp. 127–129, Kaohsiung, Taiwan, July 2005.

[11] C.-H. Wu, Y.-H. Chiu, and K.-W. Cheng, "Error-tolerant sign retrieval using visual features and maximum a posteriori estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 4, pp. 495–508, 2004.

[12] F. Quek, "Toward a vision-based hand gesture interface," in *Proceedings of the conference on Virtual reality software and technology (VRST '94)*, G. Singh, S. K. Feiner, and D. Thalmann, Eds., pp. 17–31, World Scientific, Singapore, August 1994.

[13] M. Tyrone, "Overview of capture techniques for studying sign language phonetics," in *Proceedings of the International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction (GW '01)*, pp. 101–104, London, UK, April 2001.

[14] G. Awad, J. Han, and A. Sutherland, "A unified system for segmentation and tracking of face and hands in sign language recognition," in *Proceedings of the 18th International Conference on Pattern Recognition (ICPR '06)*, vol. 1, pp. 239–242, Hong Kong, August 2006.

[15] E.-J. Holden, G. Lee, and R. Owens, "Australian sign language recognition," *Machine Vision and Applications*, vol. 16, no. 5, pp. 312–320, 2005.

[16] N. Habili, C. C. Lim, and A. Moini, "Segmentation of the face and hands in sign language video sequences using color and motion cues," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 8, pp. 1086–1097, 2004.

[17] I. Imagawa, H. Matsuo, R.-I. Taniguchi, D. Arita, S. Lu, and S. Igi, "Recognition of local features for camera-based sign language recognition system," in *Proceedings of the 15th International Conference on Pattern Recognition (ICPR '00)*, vol. 4, pp. 849–853, Barcelona, Spain, September 2000.

[18] Y. Cui and J. Weng, "A learning-based prediction-and-verification segmentation scheme for hand sign image sequence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 8, pp. 798–804, 1999.

[19] E.-J. Ong and R. Bowden, "A boosted classifier tree for hand shape detection," in *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition (FGR '04)*, pp. 889–894, Seoul, Korea, May 2004.

[20] P. Dreuw, T. Deselaers, D. Rybach, D. Keysers, and H. Ney, "Tracking using dynamic programming for appearance-based sign language recognition," in *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (AFGR '06)*, pp. 293–298, Southampton, UK, April 2006.

[21] H. Hienz and K. Grobel, "Automatic estimation of body regions from video images," in *Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction (GW '97)*, pp. 135–145, Bielefeld, Germany, September 1997.

[22] J. Wu and W. Gao, "The recognition of finger-spelling for Chinese sign language," in *Proceedings of the International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction (GW '01)*, pp. 96–100, London, UK, April 2001.

[23] V. I. Pavlovic, R. Sharma, and T. S. Huang, "Visual interpretation of hand gestures for human-computer interaction: a review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 677–695, 1997.

[24] X. Yang, F. Jiang, H. Liu, H. Yao, W. Gao, and C. Wang, "Visual sign language recognition based on HMMs and auto-regressive HMMs," in *Proceedings of the 6th International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction (GW '05)*, pp. 80–83, Berder Island, France, May 2005.

[25] R. Bowden, D. Windridge, T. Kadir, A. Zisserman, and M. Brady, "A linguistic feature vector for the visual interpretation of sign language," in *Proceedings of the 8th European Conference on Computer Vision (ECCV '04)*, pp. 390–401, Prague, Czech Republic, May 2004.

[26] C.-C. Chang and C.-M. Pengwu, "Gesture recognition approach for sign language using curvature scale space and hidden Markov model," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '04)*, vol. 2, pp. 1187–1190, Taipei, Taiwan, June 2004.

[27] M.-H. Yang, N. Ahuja, and M. Tabb, "Extraction of 2D motion trajectories and its application to hand gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1061–1074, 2002.

[28] T. Kadir, R. Bowden, E. Ong, and A. Zisserman, "Minimal training, large lexicon, unconstrained sign language recognition," in *Proceedings of the 15th British Machine Vision Conference (BMVC '04)*, Kingston, UK, September 2004.

[29] Q. Munib, M. Habeeba, B. Takruria, and H. A. Al-Malik, "American sign language (ASL) recognition based on Hough transform and neural networks," *Expert Systems with Applications*, vol. 32, no. 1, pp. 24–37, 2007.

[30] O. Al-Jarrah and A. Halawani, "Recognition of gestures in Arabic sign language using neuro-fuzzy systems," *Artificial Intelligence*, vol. 133, no. 1-2, pp. 117–138, 2001.

[31] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[32] V. I. Pavlovic, "Dynamic Bayesian networks for information fusion with applications to human-computer interfaces," Ph.D. thesis, University of Illinois at Urbana-Champaign, Champaign, III, USA, 1999.

[33] A. Bobick and J. Davis, "Real-time recognition of activity using temporal templates," in *Proceedings of the 3rd Workshop on Applications of Computer Vision (WACV '96)*, pp. 39–42, Sarasota, Fla, USA, December 1996.

[34] S. C. W. Ong, S. Ranganath, and Y. V. Venkatesh, "Understanding gestures with systematic variations in movement dynamics," *Pattern Recognition*, vol. 39, no. 9, pp. 1633–1648, 2006.

[35] H. Sagawa, M. Takeuchi, and M. Ohki, "Methods to describe and recognize sign language based on gesture components represented by symbols and numerical values," *Knowledge-Based Systems*, vol. 10, no. 5, pp. 287–294, 1998.

[36] S. C. W. Ong and S. Ranganath, "Automatic sign language analysis: a survey and the future beyond lexical meaning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 873–891, 2005.

[37] U. Zeshan, "Aspects of Türk Isaret Dili," *Sign Language & Linguistics*, vol. 6, no. 1, pp. 43–75, 2003, (Turkish).

[38] J. Ma, W. Gao, and R. Wang, "A parallel multistream model for integration of sign language recognition and lip motion," in *Proceedings of the 3rd International Conference on Advances in Multimodal Interfaces (ICMI '00)*, pp. 582–589, Beijing, China, October 2000.

[39] U. M. Erdem and S. Sclaroff, "Automatic detection of relevant head gestures in American sign language communication," in *Proceedings of the 16th International Conference on Pattern Recognition (ICPR '02)*, vol. 1, pp. 460–463, Quebec, Canada, August 2002.

[40] M. Xu, B. Raytchev, K. Sakaue, et al., "A vision-based method for recognizing non-manual information in Japanese sign language," in *Proceedings of the 3rd International Conference on Advances in Multimodal Interfaces (ICMI '00)*, pp. 572–581, Beijing, China, October 2000.

[41] K. W. Ming and S. Ranganath, "Representations for facial expressions," in *Proceedings of the 7th International Conference on Control, Automation, Robotics and Vision (ICARCV '02)*, vol. 2, pp. 716–721, Singapore, December 2002.

[42] G. Fang, W. Gao, and D. Zhao, "Large-vocabulary continuous sign language recognition based on transition-movement models," *IEEE Transactions on Systems, Man, and Cybernetics A*, vol. 37, no. 1, pp. 1–9, 2007.

[43] L. G. Zhang, X. Chen, C. Wang, Y. Chen, and W. Gao, "Recognition of sign language subwords based on boosted hidden Markov models," in *Proceedings of the 7th International Conference on Multimodal Interfaces (ICMI '05)*, pp. 282–287, Torento, Italy, October 2005.

[44] G. Fang, W. Gao, and D. Zhao, "Large vocabulary sign language recognition based on fuzzy decision trees," *IEEE Transactions on Systems, Man, and Cybernetics A*, vol. 34, no. 3, pp. 305–314, 2004.

[45] M. Zahedi, D. Keysers, and H. Ney, "Pronunciation clustering and modeling of variability for appearance-based sign language recognition," in *Proceedings of the 6th International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction (GW '05)*, pp. 68–79, Berder Island, France, May 2005.

[46] M. Sarfraz, Y. A. Syed, and M. Zeeshan, "A system for sign language recognition using fuzzy object similarity tracking," in *Proceedings of the 9th International Conference on Information Visualisation (IV '05)*, pp. 233–238, London, UK, July 2005.

[47] C. Wang, X. Chen, and W. Gao, "A comparison between etymon- and word-based Chinese sign language recognition systems," in *Proceedings of the 6th International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction (GW '05)*, pp. 84–87, Berder Island, France, May 2005.

[48] C. Vogler and D. Metaxas, "Handshapes and movements: multiple-channel American sign language recognition," in *Proceedings of the 5th International Gesture Workshop on Gesture and Sign Languages in Human-Computer Interaction (GW '04)*, pp. 247–258, Genova, Italy, April 2004.

[49] G. Fang, X. Gao, W. Gao, and Y. Chen, "A novel approach to automatically extracting basic units from Chinese sign language," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04)*, vol. 4, pp. 454–457, Cambridge, UK, August 2004.

[50] S. K. Liddell and R. E. Johnson, "American sign language: the phonological base," *Sign Language Studies*, vol. 64, pp. 195–278, 1989.

[51] R. Wilbur and A. Kak, "Purdue RVL-SLLL American sign language database," Tech. Rep. TR-06-12, School of Electrical and Computer Engineering, Purdue University, West Lafayette, Ind, USA, 2006.

[52] M. Kadous, "Temporal classification: extending the classification paradigm to multivariate time series," Ph.D. thesis, School of Computer Science and Engineering, University of New South Wales, Sydney, Australia, 2002.

[53] A. Edwards, "Progress in sign languages recognition," in *Proceedings of the International Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction (GW '97)*, pp. 13–21, Bielefeld, Germany, September 1997.

[54] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[55] E. Hjelmås and B. K. Low, "Face detection: a survey," *Computer Vision and Image Understanding*, vol. 83, no. 3, pp. 236–274, 2001.

[56] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002.

[57] Y. Tian, T. Kanade, and J. Cohn, "Facial expression analysis," in *Handbook of Face Recognition*, S. Z. Li and A. K. Jain, Eds., Springer, New York, NY, USA, 2005.

[58] B. Fröba and A. Ernst, "Face detection with the modified census transform," in *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition (FGR '04)*, pp. 91–96, Seoul, Korea, May 2004.

[59] C. Garcia and M. Delakis, "Convolutional face finder: a neural architecture for fast and robust face detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 11, pp. 1408–1423, 2004.

[60] J. Ruiz-del-Solar, R. Verschae, P. Vallejos, and M. Correa, "Face analysis for human computer interaction applications," in *Proceedings of the 2nd International Conference on Computer Vision Theory and Applications (VISAPP '07)*, Barcelona, Spain, March 2007.

[61] C. Waring and X. Liu, "Rotation invariant face detection using spectral histograms and support vector machines," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '06)*, pp. 677–680, Atlanta, Ga, USA, October 2006.

[62] B. Wu, H. Ai, C. Huang, and S. Lao, "Fast rotation invariant multi-view face detection based on real adaboost," in *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition (FGR '04)*, pp. 79–84, Seoul, Korea, May 2004.

[63] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: an overview," in *Issues in Visual and Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, Eds., MIT Press, Cambridge, Mass, USA, 2004.

[64] V. Attina, D. Beautemps, M.-A. Cathiard, and M. Odisio, "A pilot study of temporal organization in Cued Speech production of French syllables: rules for a Cued Speech synthesizer," *Speech Communication*, vol. 44, no. 1–4, pp. 197–214, 2004.

[65] N. Aboutabit, D. Beautemps, and L. Besacier, "Automatic identification of vowels in the Cued Speech context," in *Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP '07)*, Hilvarenbeek, The Netherlands, August-September 2007.

[66] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[67] D. Terzopoulos and K. Waters, "Analysis and synthesis of facial image sequences using physical and anatomical models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 6, pp. 569–579, 1993.

[68] P. S. Aleksic, J. J. Williams, Z. Wu, and A. K. Katsaggelos, "Audio-visual speech recognition using MPEG-4 compliant visual features," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1213–1227, 2002.

[69] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam, "Use of active shape models for locating structures in medical images," *Image and Vision Computing*, vol. 12, no. 6, pp. 355–365, 1994.

[70] N. Eveno, A. Caplier, and P.-Y. Coulon, "Automatic and accurate lip tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 706–715, 2004.

[71] L. Zhang, "Estimation of the mouth features using deformable templates," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '97)*, vol. 3, pp. 328–331, Santa Barbara, Calif, USA, October 1997.

[72] B. Beaumesnil, F. Luthon, and M. Chaumont, "Liptracking and MPEG4 animation with feedback control," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, vol. 2, pp. 677–680, Toulouse, France, May 2006.

[73] J. Luettin, N. A. Thacker, and S. W. Beet, "Statistical lip modeling for visual speech recognition," in *Proceedings of the 8th European Signal Processing Conference (EUSIPCO '96)*, Trieste, Italy, September 1996.

[74] P. Gacon, P.-Y. Coulon, and G. Bailly, "Nonlinear active model for mouth inner and outer contours detection," in *Proceedings of the 13th European Signal Processing Conference (EUSIPCO '05)*, Antalya, Turkey, September 2005.

[75] S. Stillittano and A. Caplier, "Inner lip segmentation by combining active contours and parametric models," in *Proceedings of the 3rd International Conference on Computer Vision Theory and Applications (VISAPP '08)*, Madeira, Portugal, January 2008.

[76] J. L. Schwartz, J. Robert-Ribes, and P. Escudier, "Ten years after summerfield: a taxonomy of models for audio-visual fusion in speech perception," in *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-Visual Speech*, pp. 85–108, Psychology Press, Hove, UK, 1998.

[77] N. Aboutabit, D. Beautemps, and L. Besacier, "Lips and hand modelling for recognition of the Cued Speech gestures: the French Vowel Case," to appear in *Speech Communication*.

[78] N. Aboutabit, D. Beautemps, and L. Besacier, "Hand and lips desynchronization analysis in French Cued Speech: automatic temporal segmentation of visual hand flow," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '06)*, Toulouse, France, May 2006.

[79] P. Taylor and A. Black, "Speech synthesis by phonological structure matching," in *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH '99)*, pp. 623–626, Budapest, Hungary, September 1999.

[80] T. Okadome, T. Kaburagi, and M. Honda, "Articulatory movement formation by kinematic triphone model," in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC '99)*, vol. 2, pp. 469–474, Tokyo, Japan, October 1999.

[81] P. Duchnowski, D. S. Lum, J. C. Krause, M. G. Sexton, M. S. Bratakos, and L. D. Braida, "Development of speechreading supplements based on automatic speech recognition," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 4, pp. 487–496, 2000.

[82] S. Minnis and A. Breen, "Modeling visual coarticulation in synthetic talking heads using a lip motion unit inventory with

concatenative synthesis," in *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP '00)*, pp. 759–762, Beijing, China, October 2000.

[83] G. Gibert, G. Bailly, D. Beautemps, F. Elisei, and R. Brun, "Analysis and synthesis of the 3D movements of the head, face and hand of a speaker using cued speech," *Journal of Acoustical Society of America*, vol. 118, no. 2, pp. 1144–1153, 2005.

[84] M. Tamura, S. Kondo, T. Masuko, and T. Kobayashi, "Text-to-audio-visual speech synthesis based on parameter generation from HMM," in *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH '99)*, pp. 959–962, Budapest, Hungary, September 1999.

[85] O. Govokhina, G. Bailly, and G. Breton, "Learning optimal audiovisual phasing for a HMM-based control model for facial animation," in *Proceedings of the 6th ISCA Workshop on Speech Synthesis (SSW '07)*, Bonn, Germany, August 2007.

[86] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: driving visual speech with audio," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '97)*, pp. 353–360, Los Angeles, Calif, USA, August 1997.

[87] E. Cosatto and H. P. Graf, "Sample-based of photo-realistic talking heads," in *Proceedings of the Computer Animation Conference (CA '98)*, pp. 103–110, Philadelphia, Pa, USA, June 1998.

[88] T. Ezzat and T. Poggio, "MikeTalk: a talking facial display based on morphing visemes," in *Proceedings of the Computer Animation Conference (CA '98)*, pp. 96–102, Philadelphia, Pa, USA, June 1998.

[89] T. Ezzat, G. Geiger, and T. Poggio, "Trainable videorealistic speech animation," *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 388–398, 2002.

[90] B. Theobald, J. Bangham, I. Matthews, and G. Cawley, "Visual speech synthesis using statistical models of shape and appearance," in *Proceedings of the Auditory-Visual Speech Processing Workshop (AVSP '01)*, pp. 78–83, Aalborg, Denmark, September 2001.

[91] D. Massaro, M. Cohen, and J. Beskow, "Developing and evaluating conversational agents," in *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Eds., pp. 287–318, MIT Press, Cambridge, Mass, USA, 2000.

[92] J. Beskow and M. Nordenberg, "Data-driven synthesis of expressive visual speech using an MPEG-4 talking head," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, pp. 793–796, Lisbon, Portugal, September 2005.

[93] E. Vatikiotis-Bateson, T. Kuratate, M. Kamachi, and H. Yehia, "Facial deformation parameters for audiovisual synthesis," in *Proceedings of the Auditory-Visual Speech Processing Conference (AVSP '99)*, pp. 118–122, Santa Cruz, Calif, USA, August 1999.

[94] G. Bailly, M. Bérar, F. Elisei, and M. Odisio, "Audiovisual speech synthesis," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 331–346, 2003.

[95] G. Geiger, T. Ezzat, and T. Poggio, "Perceptual evaluation of video-realistic speech," Tech. Rep., Massachusetts Institute of Technology, Cambridge, Mass, USA, 2003.

[96] R. Bowden, "Learning non-linear models of shape and motion," Ph.D. thesis, Department of Systems Engineering, Brunel University, London, UK, 1999.

[97] R. M. Uchanski, L. A. Delhorne, A. K. Dix, L. D. Braida, C. M. Reed, and N. I. Durlach, "Automatic speech recognition to aid the hearing impaired: prospects for the automatic generation of cued speech," *Journal of Rehabilitation Research and Development*, vol. 31, no. 1, pp. 20–41, 1994.

[98] B. Haskell and C. Swain, "Segmentation and sign language synthesis," WO 98/53438. USA, AT&T, 1998.

[99] H. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Facial animation and head motion driven by speech acoustics," in *Proceedings of the 5th Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, pp. 265–268, Kloster Seeon, Germany, May 2000.

[100] T. Toda, A. Black, and K. Tokuda, "Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis," in *Proceedings of the 5th International Speech Synthesis Workshop (ISCA '04)*, pp. 26–31, Pittsburgh, Pa, USA, June 2004.

[101] S. Hiroya and T. Mochida, "Multi-speaker articulatory trajectory formation based on speaker-independent articulatory HMMs," *Speech Communication*, vol. 48, no. 12, pp. 1677–1690, 2006.

[102] J. P. H. van Santen, L. C. W. Pols, M. Abe, D. Kahn, E. Keller, and J. Vonwiller, "Report on the third ESCA TTS workshop evaluation procedure," in *Proceedings of the 3rd ESCA/COCOSDA Workshop on Speech Synthesis*, pp. 329–332, Jenolan Caves, Australia, November 1998.

[103] P. Boula de Mareüil, C. d'Alessandro, A. Raake, G. Bailly, M.-N. Garcia, and M. Morel, "A joint intelligibility evaluation of French text-to-speech systems: the EvaSy SUS/ACR campaign," in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06)*, pp. 2034–2037, Genoa, Italy, May 2006.

[104] I. S. Pandzic, J. Ostermann, and D. Millen, "User evaluation: synthetic talking faces for interactive services," *The Visual Computer*, vol. 15, no. 7-8, pp. 330–340, 1999.

[105] G. Gibert, G. Bailly, and F. Elisei, "Evaluating a virtual speech cuer," in *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP '06)*, pp. 2430–2433, Pittsburgh, Pa, USA, September 2006.

[106] W. D. Voiers, "Evaluating processed speech using the diagnostic rhyme test," *Speech Technology*, vol. 1, no. 4, pp. 30–39, 1983.

[107] P. Boyes Braem, "Rhythmic temporal patterns in the signing of early and late learners of German Swiss Sign Language," *Language and Speech*, vol. 42, no. 2-3, pp. 177–208, 1999.