Speech technologies for augmented communication

Gérard Bailly, Pierre Badin, Denis Beautemps, Frédéric Elisei

Speech technologies for augmented communication

Gérard Bailly, Pierre Badin, Denis Beautemps & Frédéric Elisei
GIPSA-Lab, Dpt. of Speech & Cognition, CNRS/Universities of Grenoble – France

Contact author: gerard.bailly@gipsa-lab.inpg.fr

BIOs:
Gérard Bailly is a senior CNRS Research Director at the Speech and Cognition Department, GIPSA-lab, Grenoble. He is now the head of the department. He has worked in the field of speech communication for more than 25 years. He supervised 20 PhD Thesis and authored 32 journal papers and more than 200 book chapters and papers in major international conferences. He coedited "Talking Machines: Theories, Models and Designs" (Elsevier, 1992) and "Improvements in Speech Synthesis" (Wiley, 2002). He is associate editor for the Journal of Acoustics, Speech & Music Processing and reviewer for many international journals. He is a founder member of the ISCA SynSIG and SproSIG special-interest groups. His current interest is multimodal and situated interaction with conversational agents using speech, facial expressions, head movements and eye gaze.

Pierre Badin is a senior CNRS Research Director at the Speech and Cognition Department, GIPSA-lab, Grenoble. Head of the 'Vocal Tract Acoustics" team from 1990 to 2002, associate director of the Grenoble ICP from 2003 to 2006, he is adjunct to the Department head since 2007. He has worked in the field of speech communication for more than 30 years. He gained international experience through extended research periods in Sweden, Japan and UK, and is involved in a number of national and international projects. He is associate editor for speech at Acta Acustica, and reviewer for many international journals. His current interest is speech production and articulatory modelling, with an emphasis on data acquisition (MRI, ElectroMagnetoArticulograph, Aerodynamics, etc.), development of virtual talking heads for augmented speech, and speech inversion.

Frédéric Elisei is a CNRS Research Engineer at the Speech and Cognition Department, GIPSA-lab, Grenoble. He is responsible for the development and exploitation of the MICAL experimentation platform, designed to study face-to-face speech communication either between two humans or between a human and a virtual conversational agent, involving speech, eye gaze, facial expression and gestures. He works on audiovisual speech i.e. modelling and synthesis of 3D talking heads, addressing several speakers and target languages. His current interest is multimodal and situated interaction with conversational agents, in particular giving agents adaptive skills such as varying speech styles (whisper, hyper-articulation...),displaying various facial expressions or adapting the language or the phonological repertoire to the human interlocutor.

1 Introduction

Speech is very likely the most natural communication mean for humans. However, there are various situations in which audio speech cannot be used because of disabilities or adverse environmental conditions. Resorting to alternative methods such as *augmented speech* is a therefore an interesting approach. This chapter presents computer-mediated communication technologies that allow such an approach (see Figure 1). Speech of the emitter may in fact:

- not be captured by available hardware communication channels – camera, microphone
- be impoverished by the quality of the hardware or the communication channel
- be impoverished because of environmental conditions or because of motor impairments of the interlocutor.

On the reception side, Augmented Speech Communication (ASC) may also compensate for perceptual deficits of the user by enhancing the captured signals or adding multimodal redundancy by synthesizing new perceptual channels or adding new features to existing channels. In order to improve human-human communication ASC can make use of *a priori* knowledge on multimodal coherence of speech signals, user/listener voice characteristics or more general linguistic and phonological structure on the spoken language or vocabulary being exchanged. The nature of this *a priori* knowledge, the quantitative models that implement it and their capabilities to enhance the available communication signals influence the precision and robustness of the communication.

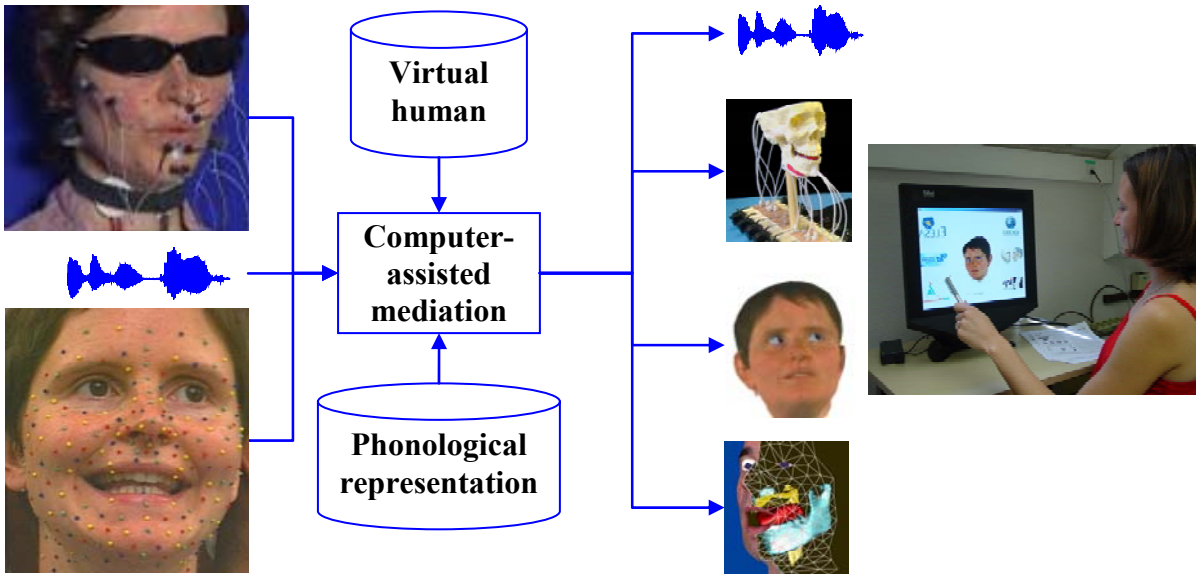

Figure 1 : Computer-mediated communication consists in driving an artificial agent from signals captured on the source speaker. The embodiment of the agent may be quite diverse: from pure audio through audiovisual rendering of speech by avatars to a more videorealistic animations by means of of virtual clones of the source speaker or anthropoid robots - here the animatronic talking head Anton developed at U. of Sheffield (Hofe & Moore, 2008). The control signals of these agents can encompass not only audible and visible consequences of articulation but also control posture, gaze, facial expressions or head/hand movements. Signals captured on the source speaker provide partial information on speech activity such as brain or muscular activity, articulatory movements, speech or even scripts produced by the source speaker. Such systems exploit *a priori* knowledge on the mapping between captured and synthesized signals labelled here as "virtual human" and "phonological representation": these resources that know about the coherence between observed and generated signals can be either statistical or procedural.

The chapter will first present:

- the signals that can characterise the speech production activity i.e. from electromagnetic signals from brain activity, through articulatory movements, to their audiovisual traces
- devices that can capture these signals with various impact on articulation and constraints on usage
- available technologies that have been proposed for mapping these various speech representations between each other i.e. virtual human, direct statistical mapping or speech technologies using a phonetic pivot obtained by speech recognition techniques

Three ASC systems developed in the MPACIF Team at GIPSA-Lab will then be described in detail:

a) a system that converts non audible murmur into audiovisual speech for silent speech communication (Tran, Bailly, & Loevenbruck, submitted; Tran, Bailly, Loevenbruck, & Toda, 2008)
b) a system that converts silent cued speech (Cornett, 1967) into audiovisual speech and vice-versa. This system aims at computer-assisted audiovisual telephony for deaf users (Aboutabit, Beautemps, & Besacier, Accepted; Beautemps et al., 2007)
c) a system that computes and displays virtual tongue movements from audiovisual input for pronunciation training (Badin, Elisei, Bailly, & Tarabalka, 2008; Badin, Tarabalka, Elisei, & Bailly, 2008).

Preliminary results of the evaluation of these three systems will be given and commented. A discussion on both scientific and technological challenges and limitations will conclude the chapter.

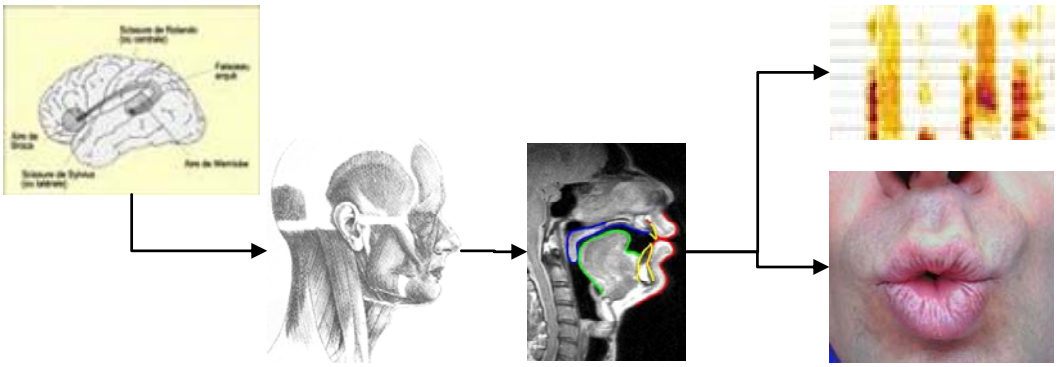

Figure 2 : The speech production chain. The intended message is decoded by the listener on the basis of audible and visible traces of speech articulation combined with *a priori* knowledge on the speaker, the language spoken and the message content given the history of the dialog and the situation.

2 Characterizing speech production

The speech production chain sketched in Figure 2 consists of several signal transformations: the electrical activity of neural circuitry drives the contraction of several dozen of muscles that further shape the geometry of the vocal tract. The air flow generated by the pressure induced by the respiratory muscles interacts with the vocal tract walls in relation with the biomechanical properties of the speech articulators and generates various acoustic sources, such as pseudo-period signals at the glottis or noise signals at constrictions. These acoustic sources excite the vocal tract resonators and are finally radiated as speech sound through the mouth, the nose and skin. Speech production can be thus characterized by:

- Neural activity. Several brain areas are activated in motor control of speech. Nota and Honda (Nota & Honda, 2004) found for example that the bilateral motor cortex and the inferior cerebellum hemisphere were activated after the subtraction for breathing, non speech vocalization, and hearing. They asked subjects to plan speech in four different conditions: A) normal speech (spoken aloud), B) mouthed speech (mouthing silently with normal breathing), C) unarticulated speech (voicing ‘‘ah…’’ without articulation), and D) internal speech. Activations were also found in the superior temporal gyrus and inferior parietal lobule of the left hemisphere. Activations are also found in Broca’s area, the supplementary motor area (SMA), or the insula, especially in case of difficult or unusual speech production task. Note also that most areas dedicated to speech perception are also activated during speech production and vice versa (Wilson, Saygin, Sereno, & Iacoboni, 2004).
- Muscular activities. Speech production involves the activation of the respiratory muscles (inhalation and exhalation), of muscles controlling the mandible, the lips, the soft palate, the pharynx and the larynx. Note also that the control of speech articulation involves the displacement of intermediate structures such as the hyoid bone. Speech production is thus accompanied by active and passive (resistive) action of both agonist and antagonist muscles.
- Vocal tract geometry. Contractions of muscles displace the above-mentioned speech articulators that shape the vocal tract. The dynamic range of this change of geometry depends on the interaction between the air flow and the articulatory movement: vocal folds typically oscillate in the range [50-400Hz], lips tongue tip or uvula oscillate at [20-30Hz] in trills, whereas the slowest articulator, the jaw, cycles at [5-6 Hz].
- Audible signals. Changes of vocal tract geometry are made audible as they change the acoustic resonances of the vocal tract and thus shape the spectrum of the speech signal that is finally radiated. The phonological structure of world’s languages is strongly conditioned by the maximal acoustic dispersion of spectral characteristics of sounds (Schwartz, Boë, & Abry, 2007; Schwartz, Boë, Vallée, & Abry, 1997)
- Visible signals. Changes of vocal tract geometry are not all visible but movements of the jaw, the lips, parts of the movement of the larynx and the tongue are available to the interlocutor in face-to-face conversation. The benefit of audiovisual integration for speech detection, binding and comprehension has been clearly evaluated since many years (Summerfield, MacLeod, McGrath, & Brooke, 1989).

Figure 3. Capturing signatures of speech-production. Left-to-right: ultrasound imaging (from (Hueber, Chollet, Denby, Dreyfus, & Stone, 2007)), electromagnetoarticulography (EMA), electromyography (EMG).

3 Capturing speech

Various devices (see Figure 3) can capture dynamic representations of the current state of the speech production system. The aim of this section is to sketch the spectrum of available technology that can be used to record useful signals characterizing articulation and phonation.

The capture of sound vibration is usually performed by distant or head-mounted microphone. An alternative has been proposed to capture sound vibration

- The stethoscopic microphone developed by Nakajima (Nakajima, Kashioka, Shikano, & Campbell, 2003) receives sound vibration through body tissue. This device is attached to the skin of the user, for instance behind the ear. The spectral bandwidth is reduced to 0-3 kHz.

The observation of visible speech is typically done using two kinds of devices:

- *Surface deformation*. 3D range data scanners deliver very precise surface geometry together with texture information (e.g. structured light, time of flight or laser-based scanner technology). Further processing is required to compensate for head movement and to parameterize this surface with a constant number of parameters.
- *Movement of fleshpoints*. Motion capture devices (photogrammetric methods with optical flow calculation or active/passive markers) deliver movement of fleshpoints. They directly parameterize the surface with a constant number of parameters.

The observation of the internal organs does not really differ from the observation of facial movements. Three kinds of characteristics are typically monitored: density maps, positions of measurement points ("fleshpoints") and biological signals. Articulatory instrumentation includes:

- Magnetic Resonance Imaging (MRI), computerised tomography (CT), cineradiography as well as Ultra Sound Imaging (Whalen et al., 2005) provide information on the density of particular atoms or molecules within a specific volume. Some systems exploit directly the density maps as direct input. A further processing stage often retrieves surface information: if a simple threshold is often sufficient to identify the geometry of vocal tract walls in MRI or CT scan images, the determination of the tongue surface in X-ray images or ultrasound images is far more complicated. The ideal simultaneous resolution in time and space needed to observe speech movements is not available yet: the relaxation time of free hydrogen nuclei in MRI does not allow temporal sampling frequencies of more than 10-20 images per second, while noise increases drastically when acquisition rate of X-ray or ultrasound imaging are increased. Note that a further processing stage is required to determine the individual outline of the various organs in the vocal tract.
- ElectroMagnetic Articulography (EMA), ElectroPalatoGraphy (EPG) or X-ray MicroBeam (XRMB) (Kiritani, 1986) provide movement or contact information for a few measurement points attached to a speech organ. Note that EMA coils and thin wires going out of the mouth as well as the EPG artificial palate may interfere with speech movements (Recasens, 2002)
- Surface or needle ElectroMyoGraphy (EMG), ElectroGlottoGraphy (EGG) or photoglotography and the various invasive systems for measuring oral or nasal airflows deliver signals that can be directly exploited for characterizing speech activity. They are however very noisy and must be cleaned via both signal processing and *a priori* knowledge

Finally neuroprosthetics and brain-to-computer interfaces (BCI) exploit devices sensitive to the electromagnetic waves created by the neurons. Invasive (brain implants), partially-invasive (Electrocorticography or ECoG) and non-invasive (electroencephalography or EEG) devices that deliver signals related to speech planning as well as loud, silent or even simulated articulation.

4 Mapping signals

ASC systems aim at restoring or even augmenting the signals characterizing articulation based on the signals that have been captured by some of the devices mentioned above. Most of these signals are noisy and deliver

incomplete information on the articulation process. The many-to-one/one-to-many mapping between these signals is underspecified and both *a priori* knowledge and regularization techniques should be used to recover the necessary information on the articulation. *A priori* knowledge can be extracted from multiple sources:

- speech maps (Abry, Badin, & Scully, 1994) that are trained off-line and memorize the possible links between these signals that represent the coherence of the speech production process. Such a system builds a kind of speech homunculus that combines all kinaesthetic and sensory-motor information collected during speech production
- the phonetic and phonological structure of the language being spoken
- … as well as higher-level information on the linguistic content of the message.

Various technological tools (Guenther, Ghosh, & Tourville, 2006; Kröger, Birkholz, Kannampuzha, & Neuschaefer-Rube, 2006; Ouni & Laprie, 2005) have been proposed to model this *a priori* knowledge. We present here two solutions: Gaussian Mixture modelling (GMM; see Toda et al. (Toda, Ohtani, & Shikano, 2006) for its application to voice conversion) and Hidden Markov modelling (HMM; see Rabiner (Rabiner, 1989) for its application to speech recognition) that have been using in the applications of ASC presented below.

4.1 Direct statistical mapping

Speech mapping consists in building a model of the sensory-motor links based on a collection of parallel recordings of multiple characteristic signals. Though in some instances signals can actually be recorded simultaneously (see for example the combination between EMA and US in (Aron, Berger, & Kerrien, 2008)), the same speech items are usually recorded in different experimental setups; resulting signals must be then post-aligned, often using the acoustic signal as common reference.

Voice conversion techniques (Toda, Black, & Tokuda, 2004; Toda & Shikano, 2005) can then be used to capture statistically significant correlations between pairs of input-output signals.

Characterizing input signals. Input feature vectors X_t are constructed by appending feature vectors from several frames around the current frame t. Data reduction techniques (principal component analysis in (Toda & Shikano, 2005) or linear Discriminant analysis in (Tran, Bailly, Loevenbruck, & Jutten, 2008)) are often used to limit the number of model parameters to determine when the training material is too limited.

Characterizing output signals. Output feature vectors $Y_t = [y_t, \Delta y_t]$ consist of static and dynamic features at frame t.

A GMM (Toda, Black, & Tokuda, 2005) is then trained for representing the joint probability density $p(X_t, Y_t|\mathbf{\Theta})$, where $\mathbf{\Theta}$ denotes a set of GMM parameters. The generation of the time sequence of the target static feature vector y from that of the source feature $X = [X_1, X_2 \ldots X_T]$ is performed so that a likelihood $L = p(Y/X,\mathbf{\Theta})$ is maximized. Note that the likelihood is represented as a function of y: the vector $Y = [Y_1, Y_2 \ldots Y_T]$ is represented as Wy, where W denotes a conversion matrix from the static feature sequence to the static and dynamic feature sequence, respectively y and Δy (Tokuda, Yoshimura, Masuko, Kobayashi, & Kitamura, 2000). Toda et al (Toda et al., 2005) have proposed an improved ML-based conversion method considering global variance (GV) of converted feature vectors that adds another term in the optimized likelihood.

The direct statistical mapping does not require any information on the phonetic content of the training data. Alignment of input and output feature vectors – if necessary – can be performed using an iterative procedure that combines Dynamic Time Warping with conversion so that prediction error diminishes as alignment and conversion improve.

The main advantage of direct statistical mapping resides in its ability to implicitly capture fine speaker-specific characteristics.

4.2 Mapping via phoneme recognition

In direct statistical mapping, the temporal structure of speech is implicitly modelled by considering (a) a sliding time window over the input frames and (b) both static and dynamic output features are combined to produce smooth and continuous parameter trajectories. Another way to account for the special temporal structure of speech is to consider that speech encodes phonological structures: in such an approach, a pivot phonetic representation that links all measurable signals is introduced.

The mapping process proceeds in two steps: a phonetic decoding using speech recognition techniques and an output trajectory formation using speech synthesis techniques. Both steps may use different mapping techniques between signals and phonemes such as classical HMM-based speech recognition combined with corpus-based synthesis.

But the recent success of HMM-based synthesis (Yamagishi, Zen, Wu, Toda, & Tokuda, 2008) opens the route to more integrated statistical approaches to "phonetic-aware" mapping systems.

The main advantage of phonetic-based mapping resides in its ability to explicitly introduce linguistic information as additional constraints in the underdetermined mapping problem. Both in the recognition and synthesis process, linguistic or even information structure may be exploited to enrich the constructed phonological

structure and restore information that could not be predicted on the sole basis of input signals e.g. melodic patterns from silent articulation as required for silent communication interfaces (Hueber et al., 2007).

5 Applications

Applications of ASC systems are numerous. Three main areas can be found in the literature: communication enhancement, aids to communication for speech impaired people, and language training.

5.1 Communication enhancement

ASC systems, when addressing communication enhancement, aim either at fusing multimodal input in order to enhance input signals or at adding extra multimodal signals for the interlocutor so as to compensate for noisy channel or noisy perceptual conditions due to the environment. Silent speech interfaces (SSI) fall into this category: SSI should enable speech communication to take place without emitting an audible acoustic signal. By acquiring sensor data from the human speech production process, an SSI computes audible – and potentially visible – signals. Both mapping approaches have been explored:

- Bu et al (Bu, Tsuji, Arita, & Ohga, 2005) generate speech signals from EMG signals recorded during silent speech articulation via an intermediate recognition of 5 vowels and the nasal sound /n/ by a hybrid ANN-HMM speech recognition system. The linguistic model has the hard job of restoring missing consonants based on phonotactic constraints of Japanese phonology. Similarly Hueber et al (Hueber et al., 2007) combine HMM-based speech recognition with corpus-based speech synthesis to generate an audible speech signal from silent articulatory gestures captured by US imaging and video.
- Conversely Toda et al (Toda & Shikano, 2005) use direct statistical mapping for converting non audible murmur captured by a stethoscopic microphone to audible speech signal.

We have recently shown that direct statistical mapping outperforms phonetic-aware HMM-based mapping and that multimodal input improves significantly the prediction (Tran, Bailly, Loevenbruck, & Jutten, 2008). A perceptual identification task performed on a very difficult vocabulary of Japanese VCV stimuli (see Figure 4) shows that listeners can retrieve from converted speech more than 70% of the phonetic contrasts whereas amplified input NAM is unintelligible.

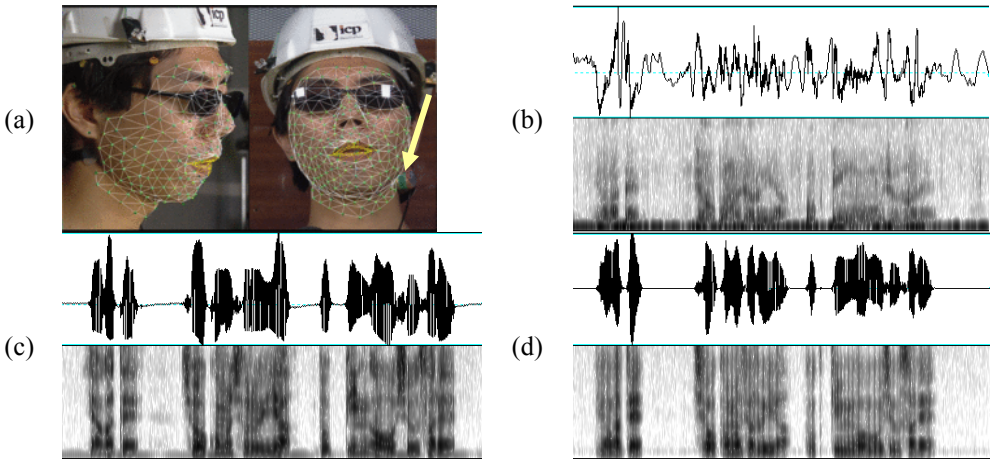

Figure 4. NAM-to-speech conversion (from (Tran, Bailly, Loevenbruck, & Jutten, 2008)). (a) 3D facial articulation tracked using an active appearance model; the position of the NAM device is indicated by an arrow; (b) non audible murmur as captured by the NAM microphone is characterized by a strong low frequency noise and a band-limited signal; (c) a target sample of the same utterance pronounced loudly in a head-set microphone; (d) the loud signal generated using GMM-based mapping from input signals (a) and (b).

5.2 Aids to communication for speech impaired people

ASC systems, when addressing communication impairment, aim to compensate for motor or perceptual deficits of one or both interlocutors. BCI can for example be exploited to offer people suffering from myopathy the ability to communicate with other people. Nakamura et al (Nakamura, Toda, Saruwatari, & Shikano, 2006) have used voice conversion of body-transmitted artificial speech to predict the structure of speech recorded before laryngectomy from speech produced after the surgery. This computer-assisted recovery of speech (Verma & Kumar, 2003) can also be performed by adapting voice fonts (Verma & Kumar, 2003) to the speaker's characteristics.

In our group, Beautemps et al (Beautemps et al., 2007) are working on a system that will enable deaf people using cued speech (CS) to have visiophonic conversations with normal hearing interlocutors. CS recognition (Aboutabit, Beautemps, Clarke, & Besacier, 2007) and synthesis (Gibert, Bailly, Beautemps, Elisei, & Brun, 2005) systems have been developed to allow conversion between speech movements and hand and lips movements. The CS-to-speech system either drives the movement of a virtual hand superposed on the video of the normal hearing speaker that produces audio speech (Bailly, Fang, Elisei, & Beautemps, 2008) or controls the movements of the face, head and hand of a virtual talking head. CS synthesis may restore more than 95% of the phonetic contrasts that could not be solved on the basis of lip reading alone (Gibert, Bailly, & Elisei, 2006).

Figure 5. Cued speech processing. Left: impressive recognition scores (Aboutabit et al., 2007) are obtained by fusing lip and hand movements. Motion capture is simplified here by make-up. Right: text-to-cued speech synthesis (Gibert et al., 2005) is performed by concatenating elementary gestural units gathered by motion capture on a human speech cuer.

5.3 Language training

Some ASC systems can also be used as tools for helping learners of a second language to master the articulation of foreign sounds. ASC systems thus perform acoustic-to-articulatory inversion: they compute the articulatory sequence that has most likely produced the sound sequence uttered by the learner. This articulation can be then displayed by means of a talking head in an augmented reality manner (see Figure 6), and compared to the required articulation so that proper corrective strategies are elicited. Several projects of virtual tutors have been launched (Engwall & Bälter, 2007; Massaro, 2006).

We have shown that despite the fact that such displays of internal articulation appear very unusual to them, listeners / viewers possess, to a certain extent, native tongue reading capabilities without intensive training (some subjects gain up to 40% recognition rate when watching the tongue display in absence of sound) (Pierre Badin et al., 2008). Such technologies may thus help people in pronunciation training.

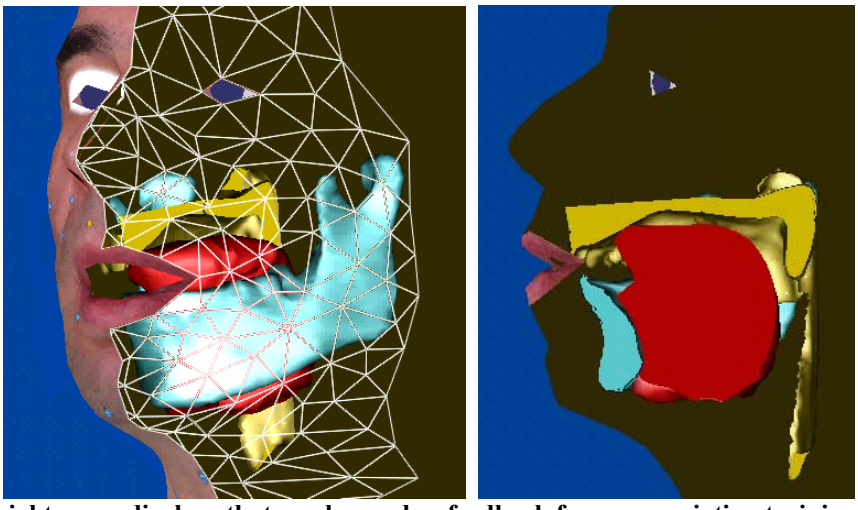

Figure 6. Artificial tongue displays that can be used as feedback for pronunciation training (from (Pierre Badin et al., 2008)).

6 Conclusions

Augmented speech communication is a very challenging research theme that requires better understanding and modelling of the speech production and perception processes. ASC systems require *a priori* knowledge to be

injected in the underdetermined inversion process so as to restore the coherence of multimodal signals that deliver incomplete information on the speech articulation or that are corrupted by noise.
A number of open issues need to be dealt with before this technology can be deployed in everyday life applications:

- The problem of speaker normalization is a hot topic: Pairs of input/output training data are only available for a limited number of subjects that have accepted to be monitored with quite invasive recording devices. To be practically acceptable, ASC systems should be able to adapt to a specific user quickly using a limited quantity of input/output data;
- Similar to speech recognition systems, ASC systems rely a lot on top-down information that constraints the mapping or inverse mapping problem. ASC should be able to benefit from language-specific constraints to gain robustness;
- Real-time issues are also very important. Guéguin et al (Guéguin, Le Bouquin-Jeannès, Gautier-Turbin, Faucon, & Barriac, 2008) have shown that full-duplex conversation is possible as long as one-way transmission delays are below 400ms. ASC systems should thus exploit limited contextual information to estimate output features. This imposes notably important constraints on speech recognition techniques;

Such technologies that connect two human brains benefit from cortical plasticity: people can learn to cope with imperfect mappings and noisy signals. Technologies that combine multimodal input and output are likely to enable computer-mediated conversation with minimum cognitive load. Evaluation issues are critical: people can cope with very crude communication channels but at the expense of the recruitment of intensive cognitive resources that may forbid any parallel activity.

Acknowledgments

Some of the PhD students of the team have largely contributed to settle this research in the laboratory: Nourredine Aboutabit and Viet-Ahn Tran are warmly thanked. We also thank our colleagues Panikos Héracleous, Hélène Loevenbruck and Christian Jutten for discussion and common work. Tomoki Toda from NAIST/Japan was very helpful. This experimental work could not have been conducted without the technical support from Christophe Savariaux and Coriandre Vilain. Part of this work has been supported by the PPF "Interactions Multimodales", PHC Sakura CASSIS, ANR Telma and Artis.

References

Aboutabit, N., Beautemps, D., Clarke, J., & Besacier, L. (2007). *A HMM recognition of consonant-vowel syllables from lip contours: the cued speech case.* Paper presented at the Interspeech, Antwerp, Belgium.

Aboutabit, N. A., Beautemps, D., & Besacier, L. (Accepted). Lips and hand modeling for recognition of the cued speech gestures: The French vowel case. *Speech Communication.*

Abry, C., Badin, P., & Scully, C. (1994). Sound-to-gesture inversion in speech: the Speech Maps approach. In K. Varghese & S. Pfleger & J. P. Lefèvre (Eds.), *Advanced speech applications* (pp. 182-196). Berlin: Springer Verlag.

Aron, M., Berger, M.-O., & Kerrien, E. (2008). *Multimodal fusion of electromagnetic, ultrasound and MRI data for building an articulatory model.* Paper presented at the International Seminar on Speech Production, Strasbourg, France.

Badin, P., Elisei, F., Bailly, G., & Tarabalka, Y. (2008). *An audiovisual talking head for augmented speech generation: Models and animations based on a real speaker's articulatory data.* Paper presented at the Conference on Articulated Motion and Deformable Objects, Mallorca, Spain.

Badin, P., Tarabalka, Y., Elisei, F., & Bailly, G. (2008). *Can you "read tongue movements"?* Paper presented at the Interspeech, Brisbane, Australia.

Bailly, G., Fang, Y., Elisei, F., & Beautemps, D. (2008). *Retargeting cued speech hand gestures for different talking heads and speakers.* Paper presented at the Auditory-Visual Speech Processing Workshop (AVSP), Tangalooma, Australia.

Beautemps, D., Girin, L., Aboutabit, N., Bailly, G., Besacier, L., Breton, G., Burger, T., Caplier, A., Cathiard, M. A., Chêne, D., Clarke, J., Elisei, F., Govokhina, O., Marthouret, M., Mancini, S., Mathieu, Y., Perret, P., Rivet, B., Sacher, P., Savariaux, C., Schmerber, S., Sérignat, J. F., Tribout, M., & Vidal, S.

(2007). TELMA: telephony for the hearing-impaired people. From models to user tests. Toulouse, France.

Bu, N., Tsuji, T., Arita, J., & Ohga, M. (2005). *Phoneme classification for speech synthesiser using differential EMG signals between muscles.* Paper presented at the IEEE Conference on Engineering in Medicine and Biology, Shanghai, China,.

Cornett, R. O. (1967). Cued Speech. *American Annals of the Deaf, 112*, 3-13.

Engwall, O., & Bälter, O. (2007). Pronunciation feedback from real and virtual language teachers. *Journal of Computer Assisted Language Learning, 20*(3), 235-262.

Gibert, G., Bailly, G., Beautemps, D., Elisei, F., & Brun, R. (2005). 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, 118*(2), 1144-1153.

Gibert, G., Bailly, G., & Elisei, F. (2006). *Evaluating a virtual speech cuer.* Paper presented at the InterSpeech, Pittsburgh, PE.

Guéguin, M., Le Bouquin-Jeannès, R., Gautier-Turbin, V., Faucon, G., & Barriac, V. (2008). On the evaluation of the conversational speech quality in telecommunications. *EURASIP Journal on Advances in Signal Processing*, Article ID 185248, 185215 pages.

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language, 96*(3), 280-301.

Hofe, R., & Moore, R. K. (2008). *AnTon: an animatronic model of a human tongue and vocal tract.* Paper presented at the Interspeech, Brisbane, Australia.

Hueber, T., Chollet, G., Denby, B., Dreyfus, G., & Stone, M. (2007). *Continuous-speech phone recognition from ultrasound and optical images of the tongue and lips.* Paper presented at the Interspeech, Antwerp, Belgium.

Kiritani, S. (1986). X-ray microbeam method for measurement of articulatory dynamics: techniques and results. *Speech Communication, 5*(2), 119-140.

Kröger, B. J., Birkholz, P., Kannampuzha, J., & Neuschaefer-Rube, C. (2006). *Modeling sensory-to-motor mappings using neural nets and a 3D articulatory speech synthesizer.* Paper presented at the Interspeech, Pittsburgh, PE.

Massaro, D. W. (2006). A computer-animated tutor for language learning: Research and applications. In P. E. Spencer & M. Marshark (Eds.), *Advances in the spoken language development of deaf and hard-of-hearing children* (pp. 212-243). New York, NY: Oxford University Press.

Nakajima, Y., Kashioka, H., Shikano, K., & Campbell, N. (2003). *Non-audible murmur recognition Input Interface using stethoscopic microphone attached to the skin.* Paper presented at the International Conference on Acoustics, Speech and Signal Processing.

Nakamura, K., Toda, T., Saruwatari, H., & Shikano, K. (2006). *Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech.* Paper presented at the InterSpeech, Pittsburgh, PE.

Nota, Y., & Honda, K. (2004). Brain regions involved in motor control of speech. *Acoustical Science and Technology, 25*(4), 286-289.

Ouni, S., & Laprie, Y. (2005). Modeling the articulatory space using a hypercube codebook for acoustic-to-articulatory inversion. *Journal of the Acoustical Society of America, 118*(1), 444-460.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *IEEE Trans. on Audio, Speech, and Language Processing, 77*, 257-286.

Recasens, D. (2002). An ema study of VCV coarticulatory direction. *Journal of the Acoustical Society of America, 111*(6), 2828-2840.

Schwartz, J. L., Boë, L. J., & Abry, C. (2007). Linking the Dispersion-Focalization Theory (DFT) and the Maximum Utilization of the Available Distinctive Features (MUAF) principle in a Perception-for-Action-Control Theory (PACT). In M. J. Solé & P. Beddor & M. Ohala (Eds.), *Experimental Approaches to Phonology* (pp. 104-124): Oxford University Press.

Schwartz, J.-L., Boë, L.-J., Vallée, N., & Abry, C. (1997). The Dispersion -Focalization Theory of vowel systems. *Journal of Phonetics, 25*, 255-286.

Summerfield, A., MacLeod, A., McGrath, M., & Brooke, M. (1989). Lips, teeth, and the benefits of lipreading. In A. W. Young & H. D. Ellis (Eds.), *Handbook of Research on Face Processing* (pp. 223-233). Amsterdam: Elsevier Science Publishers.

Toda, T., Black, A. W., & Tokuda, K. (2004). *Mapping from articulatory movements to vocal tract spectrum with Gaussian mixture model for articulatory speech synthesis.* Paper presented at the International Speech Synthesis Workshop, Pittsburgh, PA.

Toda, T., Black, A. W., & Tokuda, K. (2005). *Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter.* Paper presented at the International Conference on Acoustics, Speech and SIgnal Processing (ICASSP), Philadelphia, PE.

Toda, T., Ohtani, Y., & Shikano, K. (2006). *Eigenvoice conversion based on gaussian mixture model.* Paper presented at the InterSpeech, Pittsburgh, PE.

Toda, T., & Shikano, K. (2005). *NAM-to-Speech Conversion with Gaussian Mixture Models.* Paper presented at the InterSpeech, Lisbon - Portugal.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., & Kitamura, T. (2000). *Speech parameter generation algorithms for HMM-based speech synthesis.* Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Istanbul, Turkey.

Tran, V.-A., Bailly, G., & Loevenbruck, H. (submitted). Improvement to a NAM-captured whisper-to-speech system. *Speech Communication - special issue on Silent Speech Interfaces*.

Tran, V.-A., Bailly, G., Loevenbruck, H., & Jutten, C. (2008). *Improvement to a NAM captured whisper-to-speech system.* Paper presented at the Interspeech, Brisbane, Australia.

Tran, V.-A., Bailly, G., Loevenbruck, H., & Toda, T. (2008). *Predicting F0 and voicing from NAM-captured whispered speech.* Paper presented at the Speech Prosody, Campinas - Brazil.

Verma, A., & Kumar, A. (2003). *Modeling speaking rate for voice fonts.* Paper presented at the Eurospeech, Geneva, Switzerland.

Whalen, D. H., Iskarous, K., Tiede, M. T., Ostry, D., Lehnert-LeHoullier, H., & Hailey, D. (2005). The Haskins optically-corrected ultrasound system (HOCUS). *Journal of Speech, Language, and Hearing Research, 48*, 543-553.

Wilson, S. M., Saygin, A. P., Sereno, M. I., & Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience, 7*, 701-702.

Yamagishi, J., Zen, H., Wu, Y.-J., Toda, T., & Tokuda, K. (2008). *The HTS-2008 system: Yet another evaluation of the speaker-adaptive HMM-based speech synthesis system.* Paper presented at the Proc. Blizzard Challenge, Brisbane, Australia.