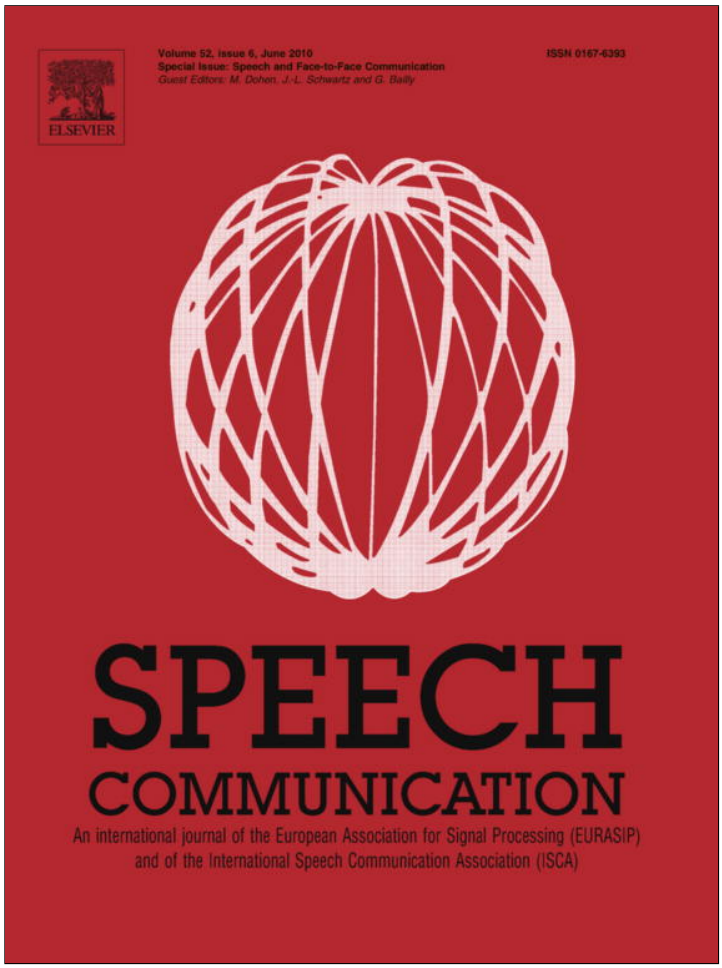

Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 52 (2010) 477–480
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Speech and face-to-face communication – An introduction

Marion Dohen *, Jean-Luc Schwartz, Gérard Bailly

GIPSA-Lab (Speech and Cognition Department, Previously ICP), UMR 5216, CNRS, Grenoble University, France

Received 26 February 2010; accepted 26 February 2010

Abstract

This issue focuses on face-to-face speech communication. Research works have demonstrated that this communicative situation is essential to language acquisition and development (e.g. naming). Face-to-face communication is in fact much more than speaking and speech is greatly influenced both in substance and content by this essential form of communication.

Face-to-face communication is multimodal: interacting involves multimodality and nonverbal communication to a large extent. Speakers not only hear but also see each other producing sounds as well as facial and more generally body gestures. Gaze together with speech contribute to maintain mutual attention and to regulate turn-taking for example. Moreover, speech communication involves not only linguistic but also psychological, affective and social aspects of interaction.

Face-to-face communication is situated: the true challenge of spoken communication is to take into account and integrate information not only from the speakers but also from the entire physical environment in which the interaction takes place. The communicative setting, the "task" in which the interlocutors are involved, their respective roles and the environmental conditions of the conversation indeed greatly influence how the spoken interaction unfolds.

The present issue aims at synthesizing the most recent developments in this topic considering its various aspects from complementary perspectives: cognitive and neurocognitive (multisensory and perceptuo-motor interactions), linguistic (dialogic face to face interactions), paralinguistic (emotions and affects, turn-taking, mutual attention), computational (animated conversational agents, multimodal interacting communication systems).

Keywords: Multimodality; Interaction; Nonverbal communication

This special issue was launched in parallel with the "speech and face to face communication" workshop organized in Grenoble in October 2008 and dedicated to the memory of Christian Benoît who died in 1998. The aim of this workshop was to show how and why speech communication must be increasingly studied in a face-to-face perspective in continuation of Christian Benoît's researches.

Speech communication is interactive. Partners involved in a spoken conversation indeed build a complex communicative act together which involves linguistic, emotional, expressive, and more generally cognitive and social dimensions. In this sense, it appears crucial to study spoken communication from an interactive point of view and the body of research on speech in interaction has recently grown actively.

This issue focuses on face-to-face speech communication. Research works have demonstrated that this communicative situation is essential to language acquisition and development (e.g. naming). Face-to-face communication is in fact much more than speaking and speech is greatly influenced both in substance and content by this essential form of communication.

Face-to-face communication is multimodal: interacting involves multimodality and nonverbal communication to a large extent. Speakers not only hear but also see each other producing sounds as well as facial and more generally

* Correspondence to: M. Dohen, INPG, 961 rue de la Houille Blanche, Domaine Universitaire, BP 46, 38402 Saint Martin d'Hères Cedex, France.
E-mail address: Marion.Dohen@gipsa-lab.grenoble-inp.fr (M. Dohen).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.02.016

body gestures. Gaze together with speech contribute to maintain mutual attention and to regulate turn-taking for example. Moreover, speech communication involves not only linguistic but also psychological, affective and social aspects of interaction.

Face-to-face communication is situated: the true challenge of spoken communication is to take into account and integrate information not only from the speakers but also from the entire physical environment in which the interaction takes place. The communicative setting, the "task" in which the interlocutors are involved, their respective roles and the environmental conditions of the conversation indeed greatly influence how the spoken interaction unfolds.

The present issue aims at synthesizing the most recent developments in this topic considering its various aspects from complementary perspectives: cognitive and neurocognitive (multisensory and perceptuo-motor interactions), linguistic (dialogic face to face interactions), paralinguistic (emotions and affects, turn-taking, mutual attention), computational (animated conversational agents, multimodal interacting communication systems). This issue focuses on a topic which has, to our knowledge, seldom been addressed in this form previously. Many related themes have been tackled, including e.g. audiovisual speech processing, dialog, multimodal communication systems, conversational agents, social robotics, and expressive communication. In this issue, we intended to focus on the convergence of these themes in a face-to-face interaction context in the speech communication area. Moreover, such an issue is timely, considering the very strong development of knowledge, techniques and systems in all the relevant areas mentioned above. We would like to dedicate this issue to the memory of Christian Benoît whose scientific achievements in the field are recalled below.

The present issue is divided into four main sections. The first section groups papers on avatars and augmented reality (Weiss et al.; Badin et al.; Heracleous et al.). Its aim is to present studies showing the potential benefits of augmented communication i.e. virtual interlocutors that supplement speech with multimodal behaviour. The second section deals with multimodal speech perception (Troille et al.; Fort et al.; Sato et al.) and presents studies investigating human–human speech communication is grounded in multimodality. The third section tackles the topic of manual gestures and prosody in communication (Colletta et al.; Cvejic et al.; Flecha-Garcia). Finally, the last section of the present issue presents studies analyzing the importance of mutual adaptation and attuning in interaction (Aubanel et al.; Kopp; Bailly et al.).

1. Avatars and augmented reality

Considering the exponential development of human–machine interaction and augmented reality systems, it appears necessary to study how such interactions are perceived by human users and how such communication could be reinforced efficiently by multimodal behaviour. The papers grouped in this section provide insight to how this can be achieved. The results from these studies could have valuable applications in crucial domains such as language tutors and speech reeducation.

An important issue in human–machine interactions and especially in the synthesis of virtual speaking avatars is their perceived visual and auditory quality. Weiss, Kühnel, Wechsung, Fagel and Möller present a study of the factors which impact the quality of talking heads in the smart home domain. They focus on the importance of voice and head characteristics as well as on interactivity and media context and compare the perceived quality of three existing talking heads. A first experiment involves non-interactive ratings of videos of talking heads on the basis of speech quality, visual quality and overall quality. A second experiment tests talking head quality in an interactive dialogue setting between user and talking head. The results show a major effect of the degree of interactivity. It appears that the interactive setting prevents the users from objectively evaluating speech and visual quality. The authors also put forward an effect of the degree of media context: distraction of the user's attention impedes their ability to objectively rate talking head quality.

Badin, Tarabalka, Elisei and Bailly question the perception of augmented speech displays. Thanks to medical imagery and articulography, it is now possible to access the very details of the speech production chain: from cortical activities to the underlying movements of the speech organs. Badin et al. address the question of how this information can be used to supplement or improve speech perception. They present a study in which they explored whether seeing the tongue could improve overall speech perception in noise. It has been extensively shown that seeing lip movements during speech facilitates speech perception in noise. The visual articulatory information provided by the lips is incomplete. Since the tongue provides crucial articulatory information, the authors developed an original tongue display system using electro-magneto-articulography (EMA) and magnetic resonance imaging (MRI) data and imagined an augmented speech perception paradigm comparing speech perception in noise in four conditions: audio only, audiovisual natural (plain face visible), audiovisual + tongue and audiovisual + jaw. They find that users can extract valuable information from tongue movements although the benefit is relatively small compared to full face display, significant only for large noise levels and after some training. Users seem to implicitly and rapidly develop 'tongue reading' abilities. These results suggest that tongue displays could be effectively useful in a number of augmented speech applications.

Providing additional information can also help in ASR especially in applications dedicated to hearing impaired subjects. In this line, Heracleous, Beautemps and Aboutabit present a study on automatic recognition of French cued speech (hand supplement providing complementary visual articulatory information for deaf people) using hid-

den Markov models (HMM). Such recognition involves lip shape as well as hand gesture recognition and a crucial aspect is the integration of the information provided by these different channels. The authors investigate phoneme and isolated word recognition and show that adding hand shape information improves overall recognition compared to the use of lip shape information only. They find similar improvements for productions of normal hearing and deaf subjects.

2. Multimodal speech perception

As mentioned above, speech is multimodal and communication involves much more than acoustic features. Speech perception in humans involves the integration of many visual cues and it is necessary to understand the underlying processes involved in order to fully comprehend how face-to-face communication functions.

Several studies have shown that the visual information could participate in speech perception processes at very early stages: speech could be seen before it is heard. Troille, Cathiard and Abry aim at exploring the timing of vowel and consonant auditory and visual streams in a CVCV context. In doing so, they find that speech can also "be heard before it is seen". They present new data aiming at accounting for these apparent contradictions. They set the perceptuo-motor link at the core of this complex phenomenology by claiming that anticipatory mechanisms in speech production set the tempo for audiovisual speech perception.

Spoken communication is also a matter of word recognition and visual information could also contribute to lexical activation processes. In order to examine this question, Fort, Spinelli, Savariaux and Kandel present a study on the word superiority effect in audiovisual speech perception. Building on the fact that seeing facial gestures enhances phonemic identification in noise, the authors investigate whether the visual information regarding consonant identification could activate lexical representations. Using a phoneme monitoring task, they find that providing visual information facilitates consonant detection in noise and accelerates the phoneme detection process. Moreover, they show a specific advantage of consonant identification in words compared to pseudo-words in the audiovisual condition, which leads them to the conclusion that visual information on phoneme identity can contribute to lexical activation processes during word recognition.

Among other considerations, the fact that the visual information, and especially the articulatory visual information, is crucial in speech perception has led to suggest that the motor cortex would be involved in speech perception: the perceiver would recognize articulatory gestures rather than or complementary to acoustic targets. Sato, Buccino, Gentilucci and Cattaneo investigate the modulation of the EMG of tongue muscles when listening to matching versus incongruent audiovisual speech stimuli. Single-pulse transcranial magnetic stimulation (TMS) was used to monitor changes in the excitability of the cortical motor representations of the tongue. They measured motor-evoked potentials in the perception of phonemes involving tongue and/or lip movements. The results show that both visual and auditory information modulate the activity in the related primary motor cortex and suggest that visual and auditory information are recoded separately in the primary motor cortex. The authors discuss these results in terms of perception/action links and theoretical modelling of audiovisual interaction.

3. Communication: prosody and manual gestures

Prosody (intonation, rhythm and phrasing) plays a crucial role in spoken communication. It has long been considered to be purely auditory/acoustic. However, a number of recent studies showed that its production and perception are actually multimodal. It remains unclear exactly which facial information cues prosody. Several studies suggest an important role of visible articulatory information, others measured correlations between eyebrow movements and fundamental frequency as well as links between rigid head motion and prosodic cues. In this line, Flecha-Garcia presents a study in which she explores the relationship between eyebrow raises and discourse structure, utterance function and pitch accents in English face-to-face dialogues. She finds that eyebrow raises occurred more frequently at the start of high-level discourse segments and in requests for or acknowledgment of information. Moreover, eyebrow raises appear to be aligned with pitch accents.

From a perceptual point of view, Cvejic, Kim and Davis present a study in which they explore the perception of prosodic contrasts from the visual information provided by the upper part of the face. Two experiments explore pairing of prosodically matching fully textured or outlined videos (no sound). Two other experiments test auditory-visual prosodic matching capabilities. The results show that participants are able to match prosody corresponding videos and sounds. The authors suggest that rigid head motion is a strong visual cue to prosody.

A growing body of research suggests that manual gestures play a very important role in spoken communication and are actually part of the communicative intention. Colletta, Pellencq and Guidetti examine age-related changes (6 year olds, 10 year olds and adults) in the production of manual gestures and speech in narratives. From transcriptions of video taped narratives, the authors find that language complexity and the use of manual gestures in communication are greatly linked to age. Deeper analyses show that co-speech gestures and their combination to speech in narratives develop with age and play a crucial role in discourse cohesion and verbal utterance framing.

4. Tuning communication in interaction

As already shown by several studies published in the present issue, studies of spoken communication are more

and more embedded in their natural context i.e. interaction. Even though this does not exclude controlled experiments to study the mechanisms involved in speech production and perception processes, it shows the necessity of considering spoken communication in an interactive framework. The papers grouped in this section put forward a number of original ideas, data and paradigms dealing with interaction in face-to-face communication.

A particular aspect of interaction is the adaptive behaviours of interlocutors involved in a conversation. In particular, speakers of different regional variations of a language manage to understand each other while identifying dialects and idiolects. Aubanel and Nguyen present a study in which they investigate the automatic recognition of phonological variation in regional accents in conversational interaction in French. They tested an automatic speech recognition technique combined to a Bayes classifier to automatically identify regional accents in natural interactive speech. Their study also confirms some of the phonetic differences between Northern and Southern French.

Further on in studying adaptive behaviours of speakers in interaction, Kopp discusses the concept of social resonance and embodied coordination in face-to-face conversation with artificial interlocutors. He analyses how and whether it would beneficial to include such considerations of adaptive behaviours of interlocutors in embodied conversational agents (ECA). He thus proposes a modelling framework both from a production and a perception point of view in a sensorimotor perspective and an implementation of this in an ECA especially in co-verbal gestures.

Many different features play a role in conversational interactions. It is necessary to understand how these features are used and which role they may play in the interaction. This will make it possible to design more natural and interactively efficient human–machine interaction systems such as ECAs. Bailly, Raidt and Elisei tackle this issue in a study on gaze behaviour of users involved in human–human and human–ECA face-to-face interactions. They describe two series of experiments in which they analyzed the interplay between speech and mutual gaze patterns during mediated face-to-face interactions. They hereby analyze both the effect of deictic gazes produced by a virtual agent on the interactant's behaviour and the interaction between cognitive state and communicative function on the measured gaze patterns.

5. *Un clin d'oeil* to Christian Benoît...

This special issue is dedicated to the memory of Christian Benoît who died in 1998. We will not recall his career here (for a biography, see e.g. http://www.icp.inpg.fr/ICP/_communication.en.html). Let us just mention that his research work intensively dealt with two of the four themes that structure the present issue: speech synthesis and audiovisual speech processing. He largely contributed to structure these domains within the speech community by organizing the first international workshops on these topics and creating the ISCA SIG groups. Even though he was less involved in the two other themes tackled in this issue, he did also work on prosody and manual gestures (through cued speech). Christian Benoît was obviously a typical face-to-face communication guy, who would certainly have entered the present field with extreme passion and enthusiasm!

Acknowledgements

The guest editors of this special issue would like to thank all the authors who submitted papers. We had a lot of pleasure reading these submissions. We also thank all the reviewers who greatly helped in the process of selecting and revising the papers submitted. Of course, we thank the editors of the Speech Communication journal who gave us the opportunity of putting together such an issue, and particularly Marc Swerts, for constant confidence and support all over the editing process.

The "Speech and Face-to-Face Communication Workshop", which launched the idea of the present issue, was organised with the support of the Christian Benoît Association (ACB: http://www.icp.inpg.fr/ICP/ACB.en.html). We acknowledge the financial support of ACB, ISCA, AFCP, GIPSA-Lab, and Mrs. Benoît.