Introduction

The books *Hearing By Eye* (Dodd and Campbell 1987) and *Speech Perception by Ear and Eye* (Massaro 1987) were the first volumes that considered speech-reading as a psychological process of interest beyond its direct applications in hearing loss and deafness (see for example, Jeffers and Barley 1971). Eight years later, David G. Stork, Christian Benoît, and N. Michael Brooke organized the landmark NATO workshop “Speechreading by Man and Machine: Models, Systems and Applications.” The workshop was “the first forum on the interdisciplinary study of speechreading (lipreading) – production, perception and learning by both humans and machines.” This workshop was followed by several volumes (Stork and Hennecke 1996; Campbell *et al.* 1998; Massaro 1998b) and was undoubtedly a major step towards the design of an audiovisual (AV) speech processing community: you will find in this volume numerous references to the series of subsequent AVSP workshops (Rhodes 1997; Terrigal 1998; Santa Cruz 1999; Scheelsminde 2001; St-Jorioz 2003; Vancouver 2005; Hilvarenbeek 2007; Tangalooma 2008; Norwich 2009; Hakone 2010) sponsored first by the European Speech Communication Association then by the AVSP Special Interest Group of the International Speech Communication Association, both bodies in which Christian Benoît was constantly promoting AV speech processing. These workshops together with dedicated workshops (see for example the AV speech recognition workshop organized in 2000 by Chalapathy *et al.*) and special sessions in international conferences have fostered the development of innumerable lines of AV speech research.

The book is divided into four main parts although most chapters address most of the questions.

The first part of the book is largely devoted to AV speech perception and to two main questions concerning human AV performance: how and where (in the brain) auditory (A) and visual (V) signals combine to access the mental lexicon. Although speech can be perceived by vision alone (i.e., via lipreading/speech-reading) and visual speech perception (Bernstein) can provide sufficient phonetic information to access the mental lexicon, talking faces constitute a major part of an infant’s perceptual experience: through the process of watching
and listening while people talk to them and point out objects of the world, infants have the opportunity to attribute semantics to the sounds they hear. Developmental studies (Burnham and Sekiyama) can thus contribute to explaining how auditory and visual information combine. Idiosyncrasies of human brain circuitry also hold clues to the evolution and development of human language, and its accessibility by eye and ear (Campbell and MacSweeney). One commonly accepted term of this intersensory integration is that the two signals carry both complementary and redundant information on the phonetic properties of the original message. The striking observation is however that the integration is something more than taking the best of both worlds and that AV perception is able to perceive properties that are carried by neither modality alone (Remez).

Some answers to this puzzle could be found in a more intimate intersensory integration at the signal level, notably that which comprises the dynamic aspects of both signals (Lander and Bruce) which are in fact audible and visible traces of the same articulatory gestures.

The second part of the book is dedicated to the production and perception of visible speech i.e. speech movements. We have access to dynamic AV information (Lander and Bruce) as consequences of the acoustic and aerodynamic consequences of the motion of speech articulators. A production-aware “grounded” perception can benefit from the availability of sensorimotor maps (which may or may not include dynamic representations) whose existence has been proved to be useful for the control of most biological movements. This intersensory integration is not only necessary for perception (and therefore comprehension) but also for movement learning and control (Cathiard et al.) Accurate descriptions and models of coordinate structures linking activations of the different speech segments are thereby necessary: for instance Beautemps, Cathiard et al. describe coordination between hand and vocal tract motions in manual cued speech.

The third part of the book presents some of the latest developments in AV speech processing by machines, particularly in AV speech recognition and synthesis (Brooke and Scott). In parallel with the development of AV research, computer-generated facial animation (Parke and Waters 1996) has attracted considerable attention and progress. Areas of application have set aside the traditional field of the animation and games industry to address more challenging applications where the metaphor of face-to-face conversation is applied to human–computer interfaces (Cassell et al. 2000). Of course, we are still a long way from building a computer which can carry on a face-to-face conversation with a human and which can pass a face-to-face Turing test, i.e. one whose computed behavior cannot be detected from natural human interaction. However, noticeable progress has been made in giving computers the “gift” of AV speech (Brooke and Scott). AV speech recognition outperforms acoustic-only speech recognition especially for degraded speech (Potamianos, Chapalatti...
et al.) while the realism of facial animation has been drastically improved by image-based speech synthesis (Ezzat et al.; Slaney and Bregler). The chapter by Massaro et al. on the Baldi talking head and its further developments for virtual speech tutoring concludes this part by reporting experimental work on augmented communication.

The fourth part focuses on the nature of the information related to oro-facial gestures (head, vocal tract, and face movements), which is necessary to enable an efficient contribution of the visual component in the audiovisual processing of speech. Bateson and Munhall’s approach is based on experimental studies of the perception of multimodal natural and synthetic stimuli in which various characteristics are either degraded or carefully preserved. Bailly, Badin et al. use a modeling approach based on a careful analysis of real speakers’ data to study the main degrees of freedom of the speech production system and their impact on the audiovisual perception of speech.

This work has been accomplished because a body of researchers is now working on the various aspects of audiovisual speech processing. Most of this synergy is due to the research field itself where the majority of the paradigms of unimodal speech research have been renewed and questioned. Part of this synergy is also due to the communicative enthusiasm of researchers such as Christian Benoit.

Scientific outcomes of multimodal speech communication studies are numerous and they cover a broad scope. We acknowledge that little is said in this book about them. Indeed, it was our decision to focus mainly on basics. However, we would like to mention one of the most exciting current outcomes: Face-to-Face speech communication. Interaction loops between production and perception of speech and gestures are at the core of this aspect of human communication, transmitting via multimodal signals parallel information about what the interlocutors say, think about what they say and how they feel when they say it. Convergence or imitation phenomena, which are at the core of L1 and L2 learning process in babies and adults, result from this interaction. Face-to-face communication studies require the integration of all the mechanisms of embodied speech production and audiovisual speech communication, and combining them with social and physical interactions between humans and between humans and their environment (see notably extended papers of presentations discussed at two workshops organized in Grenoble: Abry et al. 2009; Dohen et al. 2010). The quest for neurophysiological and behavioral correlates of these sensorimotor loops constitute an exciting research program that would certainly have attracted Christian Benoit’s attention and titillated his insatiable curiosity.
1 Three puzzles of multimodal speech perception

R. E. Remez

1.1 Introduction

Why look at the talker when you listen? This question is the straightforward practical topic of a research report at the origin of studies of multimodal speech perception (Sumby and Pollack 1954). A crucial part of the answer is furnished in this early report: Looking boosts intelligibility throughout a range of listening conditions. But despite the longstanding acknowledgment of the potential of vision in speech perception, multimodal speech perception remains a test of our theories today. The benefit of watching the talker while listening is neither well described nor well understood; this predicament is a natural consequence of a unimodal research strategy that has largely followed articulated sound into the ear. This practice has spawned a range of theoretical descriptions of the perception of speech, although it is fair to say that every account feels the weight of the multimodal problem. No hypothesis about speech perception is immune to the test, and it is a severe challenge.

Technical attention to multimodal speech perception has added to the inventory of types of multimodal integration, and the accumulated new cases are no easier to accommodate than Sumby and Pollack’s original benchmarks. This essay reviews three multimodal challenges to our understanding of speech perception, with the goal of sketching the boundaries of a unified account. The approach taken here is perceptual and general in emphasis, rather than specifically phonetic or psycholinguistic. This conceptual gambit affords an opportunity to compare speech perception to other perceptual functions, in an effort to situate speech perception as a particular functional allocation of resources drawn from a common perceptual stock. As intellectual strategy, it admits the influence of Stein and Meredith (1993), whose physiological studies of living species nonetheless describe the ancient phylogeny of multisensory unity. Of course, speech is young, certainly compared to structures of the vertebrate brainstem that compose a crucial neural constituent of multisensory convergence. It will be instructive to see how the perceptual functions accommodating speech fit within this scheme, and in order to take up the challenge, three puzzles to consider are: (1) Organization, (2) Event Perception, and (3) Experience.
Three puzzles of multimodal speech perception

1.2 Organization

To apprehend an utterance, a perceiver finds the effects of speech amid ongoing sensory flux and, informed by experience with language, resolves the linguistic message. The self-evident sufficiency of listening as a means to perceiving and understanding encourages an auditory bias in explaining perception, and from that perspective the original finding of Sumby and Pollack (1954) was unprecedented.

1.2.1 Assessing audiovisual speech transmission

Sumby and Pollack had aimed to calibrate the usefulness of viewing a talker while listening, motivated by the practical aim of devising improvements in spoken exchanges within a noisy workplace. In their test, talker and perceiver sat five feet apart, on each trial the talker producing an English spondee (for example, “cupcake,” “baseball”) chosen from a fixed list of words. The perceiver wore a headphone set through which the acoustic test items were delivered. The subject also held a copy of the word list and was asked to indicate the spoken item on each trial. Three factors were manipulated to assess the contribution of vision to speech perception. First, to estimate the baseline performance for listening, a group of participants was tested who did not look at the talker. Their performance was compared to an audiovisual condition, in which the perceivers faced the live talker, looking while listening. Second, to measure visual influence as a function of auditory resolution, the level of speech relative to noise was varied in 6 dB steps from −30 dB to 0 dB, with an additional condition in which speech was presented in the clear. Third, to estimate the effect of uncertainty, the size of the set of words from which each item was drawn was also varied, from a set of eight at the smallest to 256 at the greatest. The family of curves of the results of these tests is a thing of beauty.

Some aspects of the outcome were expectable, namely, that identification performance varied inversely with the acoustic signal-to-noise ratio (S/N) both when listening alone and when listening and looking, and, that performance was poorer the larger the set of words from which the item on each trial was chosen, true no less of the listener as of the audiovisual perceiver. The greatest contribution of vision to word identification occurred with the smallest lexical set at the lowest acoustic S/N, or, as Sumby and Pollack state it, the visual contribution to speech intelligibility increases as the speech-to-noise ratio decreases. To a first approximation, the finding is coincident with common sense, namely, that listening is ordinarily sufficient, and its insufficiency for whatever cause promotes a shift in attention to include the visual supplement, even presuming that an optic-to-phonetic projection differs hugely in dynamic from an acoustic-to-phonetic projection (Auer et al. 1997). However, a crucial analysis performed by Sumby and Pollack shows a rather different characterization of the interplay...
of visual and auditory contributions. Although the absolute visual contribution to audiovisual speech perception was greatest when the S/N was smallest, the relative contribution of vision was actually constant across a wide range of S/N. To assess this, Sumby and Pollack estimated the potential contribution of vision and the actual contribution, finding the ratio of these two estimates at each S/N. Remarkably, the constancy of the ratio shows that the relative information supplied by visual observation of the talker’s face is independent of the S/N. Their schematic illustration of this assessment of the visual contribution is shown in Figure 1.1.

If the greatest absolute benefit of audiovisual speech perception occurs when the auditory component is compromised, the findings also indicate that there is benefit of vision to intelligibility regardless of the auditory baseline. This shows that the perceptual disposition to combine the two sensory streams is robust, and
provides a glimpse of the principles of the perceptual organization of speech apprehended audiovisually. To unpack the surprise in this early study, consider the conditional dependence on vision that Sumby and Pollack imply as a default hypothesis. If speech perception relied on auditory inflow only when it yielded adequate intelligibility, visual attention to a talker would be conditional on the degree of auditory failure. Instead, the finding that visual attention persists, and contributes to speech perception over wide ranges of variation in auditory success exposes two aspects of perceptual organization.

First, audiovisual attention to a talker is primary, perhaps reflecting a natural mode of attention. Second, audiovisual attention is independent of the symbolic process that projects auditory forms into phonetic attributes. The combination of visual and auditory streams occurred regardless of the degree of success or failure in lexical identification. Two ensuing projects show that this early finding of Sumby and Pollack was genuine. They conclude that beyond the propensity of a perceiver to treat visual and auditory samples of speech as combinable by default, the audiovisual dynamic is remarkably indifferent to superordinate linguistic or symbolic properties (Remez et al. 1994).

1.2.2 The autonomy of audiovisual coherence

The audiovisual speech perception study of McGurk and MacDonald (1976) is now well known. In their project, a perceiver reported the syllable pair spoken on each trial by a recorded female talker, with auditory or auditory-visual exposure. Although the auditory contribution to perception was unequivocal when it was the sole basis for consonant identification, in combination with vision it did not dominate. Instead, subjects reported a variety of compromises or fusions between the auditory and visual streams. For understanding perceptual organization, the crucial evidence was provided by specific compromises. For example, in the instance in which the audible display conveyed [pɑ] and the visible display conveyed [kɑ], a plausible audiovisual compromise is [tɑ], preserving the audible and visible stop manner, and the audible voicelessness, and compromising on consonantal articulatory place.

Although such instances of fusion of the audible and visible consonants were disclosed by each group of subjects, listeners also reported combinations that, remarkably, were not consistent with English phonotactics. The oddity of the audiovisual combination /bdɑ/ and /bgɑ/, for instance, is its atypicality (or, perhaps, its illegality given English sonorance) in words. These reports express speech perception unbiased by its service to lexical identification and released from experience of the likely sequence of phonemes in English syllables. Once the organization of the multisensory samples of speech is determined, the functions of phonetic perception yield segmental values bound to the sensory patterns, even sequences that are inconceivable given the regular properties of
the language. It is not unreasonable to take such findings as critical counter-evidence to accounts of speech perception appealing to simple interactive activation (McClelland and Elman 1986) because this phenomenon, while exhibiting interaction of sensory modalities, denies interaction of lexical knowledge and sensory resolution in the identification of segments.

1.2.3 A unimodal parallel

We found a similar function operating in perceptual organization at a fundamental level in a unimodal case. In explaining the means by which a listener finds and follows a speech signal, whether a single talker speaks in the clear or a group of talkers yak away in a cocktail party and the listener tries to pick one out of the din, the commonplace account had been Auditory Scene Analysis (Bregman 1990). This model added a well-grounded empirical base to the proposal of Wertheimer (1938), that the starting point for the perception of objects and events, whether visible or audible, was an organized sensory field, and not simply an unaltered summary of receptor activity. Through clever tests that extrapolated Wertheimer’s principles of grouping by likeness, the auditory evidence seemed to warrant an account of perceptual organization in which elements were grouped according to their similarity, or continuity, or proximity, or their temporal coincidence. Given a welter of auditory elements in sensory flux, this account described the creation of segregated perceptual streams of like elements, binding them in a domain over which perceptual analysis then occurred (cf. Triesman 1993).

In contrast to the ideal test cases that provided empirical motivation for Auditory Scene Analysis, speech posed a recalcitrant instance with its acoustically heterogeneous signal and its extremely brief elements that fade without a trace in auditory memory within 100 ms of transduction. A perceiver who sorted a speech stream into its like elements would bind the clicks together, the whistles together, the hisses together, and the hums together, losing the natural intercalation of the acoustic elements and sacrificing the precise temporal grain of the multiple acoustic correlates of phonetic expression that arguably confers perceptual robustness (Liberman and Cooper 1972). We proposed (Remez et al. 1994) that perceptual organization of speech must, instead, be keyed to coarse-grain modulation of the spectrum, but indifferent to the short-term acoustic elements composing the stream. The constituents of a speech signal would cohere in a perceptual stream, we claimed, when a physical acoustic pattern consistent with phonologically governed articulation can be sampled by a listener despite the dissimilarities among the acoustic constituents. Clearly, the key to this kind of perceptual organization is a perceptual susceptibility to the characteristic modulations imposed by articulation on an acoustic carrier (see also Smith et al. 2002; Elliott and Theunissen 2009).
Our experiments aimed to test this conjecture by attempting to disrupt perceptual organization, in order to deduce the principle of organization from the conditions in which interference succeeded or failed. After Sumby and Pollack, and after McGurk and MacDonald, we suspected that perceptual organization was certainly subphonemic, not dependent on a lexical process, and possibly subphonetic, not dependent on segment identification. In establishing the sensory integrity of speech, listeners were acutely sensitive to the pattern of frequency variation of acoustic patterns, we found. We presented a sinewave replica of a sentence with a supernumerary tone in the frequency range of the second formant. If this tone varied over a natural extent in a natural manner it impeded the integration of the elements of a speech signal, similarly replicated in tone analogs. But, perceptual organization was not impaired by an extraneous tone when it exhibited an arbitrary pattern of variation inconsistent with vocal resonance changes (also, Roberts et al. 2010).

Whether the supernumerary tone interfered or not, our control tests showed that it did not evoke phonetic impressions, a crucial bit of evidence that it was interfering with the step at which the auditory flux is resolved into streams of common origin, rather than the step at which segmental phonetic attributes are resolved. Although this finding is incompatible with a version of pandemonium proposed in a peremptory account by Liberman (1979), as an instance of perceptual verticality, Liberman (1996) conceded that it is a legitimate alternative to Auditory Scene Analysis.

### 1.2.4 The puzzle

Together, these projects allow a sketch of the characteristics of audiovisual perceptual organization preliminary to new research. For one, audiovisual attention to speech is ineluctable, as Sumby and Pollack showed, and perceivers combine visible and audible samples of speech naturally. For another, intersensory integration in perceiving speech is fundamental to success in extracting the linguistic message, but it is not contingent on the likely or regular properties. An organized stream of visible and audible samples, once formed, is analyzed as if its sensory pattern is projected into a phonetic sequence without the influence of the lexical or indeed the phonemic experience of the perceiver. By this assumption we can reconcile the reports of McGurk and MacDonald with those of Sumby and Pollack. Last, our studies with sinewave replicas of speech show that even within the auditory modality a listener does not find the speech signal by applying a standard of similarity to the variety of auditory impressions in any moment or brief temporal span. Instead, a rather abstract sensitivity to vocal modulation, independent of the elements, appears to do the trick, and such sensitivity surely applies to the intermodal circumstance. Perhaps a voiced stop hold, a release burst, and a voiced oral resonance are less dissimilar to one
another than any of these is to a 2½-D sketch of the vermilion border of the lips. But, the principle that this research defends is that of organization by sensitivity to characteristic change, rather than to characteristic elements, and is a candidate for evaluation as a principle of multimodal organization in speech perception, not just within the auditory modality.

1.3 Event perception and speech perception

A second puzzle about multimodal speech perception stems from the rather different ontology that results from introducing the visual modality to the description. As long as the theoretician’s goal is restricted to explaining the speech chain in auditory terms, the present simplifying assumptions typical of our field can endure. Specifically, the perceiver can be characterized as a listener to language, as if the expression of speech were brought about by linguistic plans alone. Similarly, the precise situation of speech within the plenum of events that engages a perceiver need not be specified. It is enough to conceptualize the listener identifying the next item in a list of syllables or words or sentences, and distinguishing the present one from all other possibilities that a vocal tract produces. In this constrained world, speech originates in no particular place, and the talker remains occult, literally, though the effects of speech are detectable by ear. The circumstances differ in an audiovisual setting.

The events of audiovisual speech perception, while arrayed through the artifice of video technology, are less readily subject to the common explanatory idealizations of auditory non-linearity or vocal tract gestures. A specific talker is finally visible, and many of the depicted characteristics are unassailable. Clearly, the theoretical burden alters, for the visible and audible talker is encountered within an ongoing scene in which durable objects and events, albeit non-distinctive linguistically, are concurrent with the acts of language production. The coalescence of event perception and speech perception is a fundamental topic of investigation now, and the basic problems are beginning to take shape. Specifically, the perception of speech appears to use different criteria, and possibly a different grain of sensitivity than the perception of events more generally.

1.3.1 Temporal coincidence and phonetic perception

Is it tautology to assert that the perceiver registers that an event includes speech, and therefore, when it is possible, perceives the linguistic properties of the speech? In a vulgar ecology of language, it must be true that the distribution of phonetic properties lies within the distribution of vocally produced sound, and it is reasonable to suppose that the perception of speech is contingent on the perception of vocalization. The relation of the two distributions is that of subset