The Oxford Handbook of
Psycholinguistics

Edited by

M. Gareth Gaskell

Consulting editors

**Gerry Altmann, Paul Bloom,
Alfonso Caramazza,
and Pim Levelt**

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

British Library Cataloguing in Publication Data

Data available

Library of Congress Cataloging-in-Publication Data

The oxford handbook of psycholinguistics / edited by M. Gareth Gaskell; consulting
editors, Gerry Altmann . . . [et al.].
P. ; cm.
Includes bibliographical references and index.
ISBN 978-0-19-856897-1 (alk. paper)
1. Psycholinguistics -- Handbooks, manuals, etc. I. Gaskell, M. Gareth. II. Title
Handbook of psycholinguistics.
[DNLM: 1. Psycholinguistics -- Handbooks. 2. Language Development -- Handbooks. BI
455.A3 098 2007]
BF455.094 2007
401'. 9 -- dc22

2007012019

Typeset in Minion
by Cepha Imaging Pvt Ltd, Bangalore, India
Printed in Great Britain
on acid-free paper by
Biddles Ltd., King's Lynn, Norfolk

ISBN 978-0-19-856897-1

10 9 8 7 6 5 4 3 2 1

CHAPTER 2

Audiovisual speech perception and word recognition

Dominic W. Massaro and Alexandra Jesse[1]

2.1 Introductlon

The goal of this chapter is to describe how our understanding of speech benefits from having the speaker's face present, and how this benefit makes transparent the nature of speech perception and word recognition. When observing modern life with the omnipresence of mobile phones, voice messaging, and streaming over the internet using VOIP, one might erroneously think of speech communication as becoming a purely auditory phenomenon. Although speaker and listener still often face each other in situations in which communication is not aided by technology, modern technology freed us from the need to talk to each other in person. Certainly, these modern communication methods find a wide acceptance, but people are reluctant to forfeit face-to-face communication.

The preference for face-to-face communication might have little to do with language understanding but could simply reflect a preference for direct human contact. However, there is evidence that our preference for talking to each other face-to-face is not only for this social norm but actually serves the purpose of providing information that aids understanding the communicated message. The face in communication is valuable for several reasons: emotion is better understood with the face presented along with the voice (Ellison and Massaro, 1997; de Gelder and Vroomen 2000; Massaro and Egan, 1996; Vroomen et al., 2001); many back-channeling and turn-taking cues essential for effective and efficient dialog are apparent in the face, gestures, and body; and of course the face adds to the intelligibility of the conversation (see Massaro, 1998 for an overview). Thus, face-to-face communication is the ideal venue for seamless exchanges among interlocutors.

If visual speech in communication is available, then an important question is if and when it is used as source of information for audiovisual speech recognition. Proponents of so-called auditory dominance (Sekiyama and Tohkura, 1991; 1993) have argued that visual speech merely is a back-up source when the auditory signal is not sufficient for recognition. However, this notion has been falsified by research showing that information from the face is used whenever available, even when the auditory signal itself is not ambiguous (McGurk and MacDonald, 1976) and when participants are instructed to ignore all visual information in their judgment (Massaro, 1987) or are instructed to simply report what they heard (McGurk and MacDonald, 1976). Furthermore, the information provided by the visual signal is not completely redundant, because the addition of visual information improves spoken word recognition above and beyond the

[1] Alexandra Jesse is now at the Max Planck Institute for Psycholinguistics, 6500 AH Nijmegen, The Netherlands.

level of performance predicted by pure redundancy (Massaro, 1998: ch. 14). People generally benefit from having visual speech available (Jesse et al., 2000/2001; MacLeod and Summerfield, 1987; Massaro and Bosseler, 2003; Sumby and Pollack, 1954), not just those with hearing loss. And as noted by Summerfield (1987: 3), bimodal speech perception "provides an opportunity to study the perception and memory of speech through a novel modality, and audio-visual speech perception provides the opportunity to study two perceptual systems working in collaboration to analyze the phonetic events." Therefore, an account of spoken word recognition needs to consider the role of visual information in speech communication.

This chapter will give an overview of the main research questions and findings unique to audiovisual speech perception research as well as discussing what general questions about speech perception and cognition the research in this field can answer. The influence of a second perceptual source in audiovisual speech perception compared to auditory speech perception immediately necessitates the question of how the information from the different perceptual sources is used to reach the best overall decision. This need to process multiple sources of information also exists in auditory speech perception, however. For example, as described in the chapter by McQueen (Chapter 3, this volume), acoustic cues and context information are naturally combined to determine the overall percept. Audiovisual speech simply shifts the focus from intramodal to intermodal sources. As we will see in section 2.3, these two forms of processing are not necessarily qualitatively different from each other. It is essential, however, that a model of speech perception operationalizes the concept of processing multiple sources of information so that quantitative predictions can be made. The main theoretical approaches to explain integration and audiovisual speech perception are introduced and critically discussed. Furthermore, this chapter provides an overview of the role of visual speech as a language learning tool in multimodal training.

2.2 Information and information processing in audiovisual speech perception

The most basic finding in research on audiovisual speech perception is that adding visual speech to auditory speech improves performance substantially. The audiovisual benefit, for example for bisyllabic words, is comparable to a 15 dB change in signal-to-noise ratio (Sumby and Pollack, 1954; as reported in Grant and Seitz, 2000). Some researchers (Grant and Seitz, 2000) hypothesize that this benefit could be even larger for continuous speech material, since visual speech could be informative about word boundaries, prosody, and stress patterns. Indeed, the correlation between visual-only recognition and audiovisual recognition scores is higher for sentences than for single consonants (Grant et al., 1998). The audiovisual benefit has been shown for speech items ranging from single syllables (e.g. Massaro et al., 1993a) to words (e.g. de la Vaux and Massaro, 2004; Sumby and Pollack, 1954), sentences (Jesse et al., 2000/2001; MacLeod and Summerfield, 1987), and even paragraphs (Reisberg et al., 1987). Although this benefit can be most easily observed when auditory speech recognition is impaired by noise, even intact auditory speech benefits from additional visual speech information (Arnold and Hill, 2001; Reisberg et al., 1987).

There is no doubt that, in general, auditory speech is more informative than visible speech. But the audiovisual recognition benefit emerges from both the complementary and redundant nature of visual and auditory speech information (see e.g. Walden et al., 1974). Auditory-visual speech complementarity means that one modality is more informative on those dimensions on which the other modality is less informative. For example, information about the manner of articulation (e.g. /ba/ vs. /ma/) and about voicing (e.g. /pa/ vs. /ba/) is easier to distinguish acoustically than visually (Massaro, 1987; 1998; Summerfield, 1987). Voicing information is fairly robust in the auditory signal even if noise is added, whereas little voicing information can be found in the visual signal. On the other hand, information about the place of articulation is highly confusable in auditory speech (e.g. /ma/ vs. /na/; Miller and Nicely, 1955), but not very confusable in visual speech. In addition, auditory place information is particularly vulnerable to the addition of auditory noise (Miller and Nicely, 1955). As a consequence of this complementarity, we would expect a lower audiovisual benefit if the response alternatives in a study share the same place of articulation than if they are relatively distinguishable on the basis of the place of articulation. This audiovisual benefit should be especially easier to show with noise added to the auditory speech signal.

There is also evidence that the two perceptual sources of information even provide complementary information about different subsets of a linguistic feature: The auditory signal seems to provide mostly information about the place of articulation for middle or back consonants, whereas the visual signal is mostly informative about the place of articulation of labial consonants (Jesse, 2005). Psychoacoustically, an alveolar segment like /d/ is highly discriminable because there may be less upward masking of the second formant by the first (Tillmann, 1985, pers. comm.), whereas a labial segment like /b/ is similar to /v/ and /ð/. Visually, the open mouth for /d/ is less prominent than the labial closure for /b/; furthermore, there is less uncertainty for /b/ because there are many more segments articulated with an open mouth than with a labial closure.

In addition to the complementarity of auditory and visual speech, the audiovisual benefit also arises from the redundancy of visual and auditory speech (Walden et al., 1974). Two redundant observations are always beneficial relative to just one if they are analyzed in the appropriate way. For example, although place information is available through visual information about mouth closure, place information is also provided by the formant structure in the auditory signal. Summerfield (1987) observes the existence of unique multimodal cues as a third reason for the existence of an audiovisual benefit. For example, in order to detect if a plosive is voiced or voiceless, the time between seeing the release of the plosive and hearing the onset of the vocal cord vibration is informative about perceiving voicing onset. On the other hand, either of these properties independently is insufficient to make a reliable voicing categorization. This audiovisual benefit arises from the time between the onset of the auditory information about the vocal cord vibration and the visual information of seeing the release of the stop consonant. Breeuwer and Plomp (1986) showed that for plosives in a vowel–consonant–vowel context the voicing feature is more accurately recognized when bimodal rather than unimodal information is provided. The presented auditory information was a sequence of glottal pulses that can be assumed to provide no information by itself about the identity of the presented consonant. The visual condition alone showed only poor performance levels that were near chance. However, adding the pulses to the visual presentation of the speaker significantly improved recognition.

The size of the audiovisual benefit therefore depends on the distribution of information within and between these two modalities, or, more specifically, on the degree of redundancy, complementarity, and audiovisual uniqueness. However, the audiovisual benefit depends not only on the available *information*, i.e. the decrease in ambiguity through the signal presentation about the nature of the percept (Shannon, 1948), but also on the *processing* of the information (Massaro and Cohen, 1999). Perceivers differ in their ability to recognize unimodal auditory and unimodal visual speech, so that these performance levels have to be taken into account when assessing if they also differ in their skill in integrating both sources of information in audiovisual speech. The question of whether or not all the information available to the perceiver is integrated can only be answered if we also know the information that is available to the perceiver. In addition, we would need to know if a poor result is due to poor integration or to other processing factors, such as limited working memory capacity or difficulties in application of linguistic knowledge, or if it is due to less auditory and/or visual information (Grant et al., 1998).

Generally, perceivers are fairly good at extracting and using information from the face. This is a robust phenomenon that can be found even when visual information is degraded, as is the case when the perceiver is not directly facing the speaker, the facial image is blurred, or is viewed from a large distance (Campbell and Massaro, 1997; Erber 1971; 1974; Jordan and Sergeant, 2000; MacDonald et al., 2000; Massaro, 1998; Munhall et al., 2004; Vitkovich and Barber, 1994). The pervasiveness and obligatory nature of the integration of visual and auditory speech was revealed by studies that showed that integration occurs even when the perceiver is instructed to ignore either the auditory or the visual information (Massaro, 1987); or if the visual and auditory information is temporally misaligned (Campbell and Dodd, 1980; Massaro and Cohen, 1993a; Massaro et al., 1996; Munhall et al., 1996; van Wassenhove, 2004).

The perceiver also integrates auditory and visual speech that is phonetically mismatched (McGurk and MacDonald, 1976). This phenomenon has been widely studied, mostly with the aid of the McGurk effect. In the McGurk illusion, an auditory /ba/ is presented together with a mismatching visual /ga/. The perceiver, however, will often have the illusion of perceiving a /da/, /va/, or /ða/ (Massaro, 1998). People are not always consciously aware of the mismatch, and will fall victim to this illusion even when instructed to ignore the lips (Summerfield and McGrath, 1984), when the gender of voice and face mismatch (Green et al., 1991), or even when silently producing one sound segment while listening to someone else's mismatching voice over headphones (Sams et al., 2005). A simplified explanation of the McGurk fusion is that the alveolar /d/ is a compromise of the contradicting place information from the bilabial /b/ and the velar /g/ (however, see Massaro, 1998).

The McGurk effect has been extensively studied (see e.g. Green and Gerdman, 1995; Green et al., 1991; MacDonald and McGurk, 1978; Mills and Thiem, 1980; Rosenblum et al., 1997; Sams et al., 1997; Sekiyama and Tohkura, 1991; 1993). The McGurk illusion gives also insight concerning the distribution of information over time. Other than fusion responses (e.g. /da/), the combination response /bda/ is also common when an auditory /da/ is presented with a visual /ba/. However, the mismatching information of a visual /ba/ and an auditory /da/ is not usually perceived as the combination /dba/. This might be because place information about /ba/ is available earlier in the visual than in the auditory signal (Massaro and Cohen, 1993a; Smeele, 1994). However, it is important to note that data obtained from the McGurk effect is not usually in an informative range to evaluate models of integration (Massaro, 2003; Schwartz, 2003). A more informative paradigm to evaluate the validity of quantitative models of integration is provided by experiments that create ambiguity by creating a continuum between two endpoint speech segments *within* a modality which are then presented as orthogonal audiovisual combinations. An expanded factorial design (Massaro, 1998) includes unimodal conditions in addition to these audiovisual combinations.

Other questions regarding integration are *when* the visual and auditory speech information is integrated and what the window of integration is. A late integration process would first categorize the information provided by each perceptual source before integration occurs. In comparison, an early integration process would integrate the support provided by the two sources of information before recognition occurs. This mechanism would integrate continuous values rather than categorical labels. Behavioral and neuropsychological research provides direct evidence for early integration (Besle et al., 2004; Colin et al., 2002; Schwartz et al., 2004) and also shows that processing models with an early integration mechanism, like the Fuzzy Logical Model of Perception (Massaro, 1998; see section 2.3 below) and the Pre-labeling Model (Braida, 1991), give a better description of audiovisual consonant recognition data than a Post-labeling Model, where only discrete labels are integrated (Braida, 1991).

One of the major challenges in speech recognition is that perceptual information about a single segment unfolds over time rather than being presented instantaneously. Not only are cues for the identity of the speech segment distributed over time, but also there are some dynamic cues in the speech signal (Kewley-Port, 1983; Rosenblum et al., 1996). Information is not packed and aligned like beads on a string. It might be believed that the challenge for the system is to determine what information belongs to which segment in time, and to integrate these pieces of information to recognize speech. In this case, the system has to keep track of all this information for a short period of time and integrate the information belonging to the same percept. This would not be a trivial task, because information given at a certain point in time can be informative about different parts of the utterance. To complicate this issue further, information belonging to the same segment is not always exactly co-occurring in time in the two modalities in audiovisual speech (Munhall and Tohkura, 1998). Often visual information is available earlier than auditory information for a certain segment (Massaro and Cohen, 1993a; Smeele, 1994). For example, if the initial phoneme of an utterance is a bilabial plosive, then the production of a bilabial closure is needed to build up air pressure for the sudden release of the plosive sound. During the production of this bilabial closure, the available visual information (i.e. closing of the lips) is accompanied by silence. Visual information about the place of articulation is available before auditory place of articulation information. Place of articulation is not the only information that is available earlier in the visual signal. Information for the identification of vowels is available earlier in the visual than in the auditory signal (Cathiard et al., 1995).

However, a less intelligent system might be equally successful in combining multiple sources of information in speech perception. Rather than having to keep track of which cues belong to which segments, the system could simply integrate the simultaneously arriving cues to obtain degrees of support for the various segments. A segment that receives successive support across time would emerge as a percept. This interpretation appears to be consistent with the observations that bimodal speech perception is relatively robust with respect to temporal asynchrony between the presentation of information from the two modalities, especially when the visual information leads (Campbell and Dodd, 1980; Massaro and Cohen, 1993a; Massaro et al., 1996; Munhall et al., 1996; van Wassenhove, 2004). Accurate recognition performance persists across small asynchronies and actually improves compared to the aligned condition when visual information leads by about 80 to 120ms (Grant and Greenberg, 2001; Greenberg and Arai, 2004). This might be due to a constraint of the earlier arriving information on the processing of later arriving information. For example, early arriving visual place of articulation information might "prime" (Greenberg and Arai, 2004: 1068) speech representations that share this place of articulation. Visual information would constrain the set of possible word alternatives before auditory information becomes available. For example, if a word like /pin/ is presented with leading visual information, then information about the labial closure is already available before the onset of auditory speech. Visually, labial closure provides information about place and manner of articulation. Therefore, early in processing the candidate set can be restricted to words starting with /b/, /m/, or /p/. Candidates like SIN, FIN, THIN, SHIN, GIN, WIN, TIN, DIN, or KIN could be excluded. As we will see in the next section on word recognition, visual speech information is indeed influencing the set of competing lexical candidates in audiovisual word recognition. By definition, this would have to occur if people successfully lipread the given visual information. The recognition advantage found for leading visual speech (80–120ms) could also be due to increased processing time required by the visual signal relative to that required by the auditory signal. For example, optimal auditory-visual localization occurs when the visual speech is presented 80–120ms before the auditory speech signal (Lewald and Guski, 2003).

2.2.1 The lexicon and word recognition

As we have seen in the previous sections, visual information is not only used whenever available in order to understand spoken language, but also contributes substantially to our understanding. It is therefore self-evident that a full account of spoken word recognition should also consider the role of visual speech information.

While the speech signal unfolds, the incoming information is evaluated in terms of its support for lexical representations (see e.g. Connine et al., 1993; Marslen-Wilson and Zwitserlood, 1989; Shillcock, 1990; Tabossi et al., 1995; Zwitserlood, 1989). The support, and therefore the recognition, of a word is dependent on the degree to which the perceptual input matches the lexical representation of the word and mismatches alternative word candidates (see Gaskell, Chapter 4 this volume; McQueen, Chapter 3 this volume). The contribution of visual speech to audiovisual word recognition should be substantial, since visual speech primarily provides information on the place of articulation of consonants, which is a critical feature in spoken word recognition (Greenberg, 2005). Many words are only distinguishable by the place of articulation of one of their constituent consonants (e.g. met vs. net; Greenberg, 2005). Visual speech should therefore be an important source for lexical recognition.

While the importance of lexical similarity has been widely studied for the recognition of auditory spoken words (see Gaskell, Chapter 4 this volume; McQueen, Chapter 3 this volume; Pisoni and Levi, Chapter 1 this volume), it has only been recently addressed in the recognition of visual spoken words. As in auditory speech, Auer and colleagues (2002; Mattys et al., 2002) found that in a speech-reading task, words with more visually highly confusable competitors are recognized less accurately than words with fewer visual highly confusable competitors. Auer's results were obtained for both deaf and normally hearing adults when presented only with visual speech. This generalizability across participants suggests that the experiment taps into a general process of speech perception. Auditorily based confusability of the target words with competitors could not account for these results. There was no significant correlation between this measure and accuracy for either participant group. This seems to be evidence that the evaluation of the auditory and visual signal are independent processes. It is also in accordance with recent results that the similarity among visual and auditory words is primarily determined by different features (e.g. Jesse, 2005). Therefore, we expect that visual and auditory competitor measures would not strongly correlate.

Auer (2002) failed, however, to find an effect of intelligibility of the segments of a target word. Keeping competitor structure and word frequency constant, a difference in segmental intelligibility of the target word produced only weak effects. Auer argued that this might be due to a restriction of variability in visual confusability. Mattys et al. (2002), on the other hand, presented evidence for the influence of intelligibility of a target word's segments on the word's correct identification by varying the target's lexical equivalence class as a measure of competitor similarity influence (Mattys et al. 2002). A lexical equivalence class (LEC) is a set of words that are visually highly confusable, if not indistinguishable. The LECs are formed on the basis of phoneme equivalence classes (PECs; Auer and Bernstein, 1997). All visual phonemes that are highly confusable with each other are defined as belonging to the same phoneme equivalence class. Or in other words, all highly confusable visual phonemes are members of the same *viseme* class (Fisher, 1968: 800). The cut-off point here is arbitrarily chosen so that 75 percent of all confusion responses to a phoneme fall within its PEC. The LEC of a word has less uncertainty than can be estimated from the PECs of a word. This is the case because not all members of one PEC are permissible in a word context. For example, if a word has a phoneme from a PEC of three phonemes, but only two out of the three phonemes form words, the lexical constraints reduce the uncertainty about this phoneme. More specifically, if for example, /m/, /b/, and /p/ are the only members of the same PEC, then given the word context /_ɪn/, /p/ and /b/ form words (*pin*, *bin*), but not /m/ (*min*). Therefore, *pin* and *bin* would still be in the same LEC, but not *min* Given that there might be a reduction in uncertainty for other phonemes of the same word as well, lexical permissibility constraints might greatly reduce the difficulty in identifying a word. Evidence for this distinction also comes from studies in auditory word recognition (Ganong, 1980; Marslen-Wilson et al., 1996) where the lexical competitor space influenced the interpretation of an ambiguous phoneme. An ambiguous sound was treated as being an example of the target word, if there was no competitor. For example, if the rhyme /_ɪn/ is preceded by a phoneme that is ambiguous between /b/ and /m/, then it will be interpreted as /b/, since only /b/ forms a word in this context. If there was a competitor that was equally supported by the auditory signal, then both competitor and target were maintained as possible candidates (Marslen-Wilson et al., 1996).

Mattys and colleagues (2002) showed for visual speech that the size of a LEC predicts identification accuracy. Words in larger LECs are less accurately recognized than words in smaller LECs. In addition, it was found that the frequency of the target word facilitated its correct identification. This was true for all LEC sizes and despite equating of the mean frequency of all LECs. Further, the effect of LEC size and frequency separately accounted for variance above and beyond the variance explained by differences in PECs.

This research on lexical similarity effects in visual speech shows that similar if not the same processes are involved in visual and auditory word recognition. For both, word recognition is determined by competition in the lexicon. This competition is influenced not only by the similarity of the input with the candidate words, but also by what phonemes are lexically permissible in a certain context. This has also been shown to be the case for audiovisual spoken word recognition. The work of Brancazio (1999; 2004) showed that the lexical status of unimodal items influences the identification of audiovisually mismatching McGurk stimuli. Visual information had the strongest influence on the percept (i.e. was most likely to lead to a fusion response) when the auditory stimulus was a non-word but the audiovisual fusion was a word. When both auditory stimulus and audiovisual fusion percept were words or non-words, the visual influence was weaker. However, the visual influence was the weakest when only the auditory stimulus was a word. This can be explained in terms of overall goodness of support (e.g. as predicted by the Fuzzy Logical Model of Perception; Massaro, 1998). For example, an auditory non-word only yields perceptual support, and not much lexical support. The visual information supports the visual target word but also words and nonwords that are similar. The fusion response is the candidate that is most supported by the information integrated from both perceptual sources. Candidates that are also words are further supported by lexical information. Therefore, the visual target will benefit not only from being the best candidate given the mismatching perceptual sources but also from the lexical support.

2.2.2 Facial information for speech perception

Visual speech contributes to successful speech recognition; however, the question arises as to what parts of the face are responsible for this phenomenon. Obviously the movements of the lips are one of the main contributors (see e.g. Ijsseldijk, 1992; Marassa and Lansing, 1995; Ouni et al., 2005; Thomas and Jordan, 2004), and are sufficient to provide an audiovisual speech advantage. In comparison, lines shaped like a mouth are not able to improve speech recognition (Summerfield, 1979). The characteristic position and movement of the lips contribute to a general audiovisual recognition benefit probably because they mainly convey visual place of articulation information. Place of articulation information can also easily be observed by position and movement of teeth and tongue (MacLeod and Summerfield, 1987; Summerfield, 1987). Information about a front place of articulation is readily apparent in visual speech (see e.g. Jesse, 2005). Labial closure, for example, indicates the production of bilabial consonants. A tuck of the lower lip underneath the upper front teeth reveals the production of labiodental fricatives. In addition to information from the lips, the visibility of teeth and tongue positions and movements during production also contribute to speech reading (Preminger et al., 1998). Consonants with a mid place of articulation (dental to postalveolar) can be distinguished by their characteristic tongue tip movement during their production. Dental consonants can further be distinguished from other consonants with a mid place of articulation by dental adduction—i.e. whether or not teeth are seen and move vertically closer during their production. Back place of articulation is less easily transmitted visually. But for example, the time-course of lip protrusion that is distinctive for /w/ helps its recognition (Summerfield, 1987). Voicing and nasality are produced by the vocal cords and the soft palate respectively, which is difficult to observe directly (however, see discussion of multimodal voicing cues above; Breeuwer and Plomp, 1986).

Vowels are discriminated visually most successfully based on the lip rounding (i.e. protrusion and opening of the lips) and the height of the tongue (Breeuwer and Plomp, 1986; Jackson et al., 1976; Montgomery and Jackson, 1983). The visual signal seems to contain more information on lip rounding than on tongue height (Benguerel and Pichora-Fuller, 1982). Diphthongs are generally more identifiable visually than monophthong vowels (Wozniak and Jackson, 1979).

Other areas in the face besides the lips are informative about visual speech, such as the movements of the jaw (Benoît et al., 1996) and the cheek (Preminger et al., 1998). There is information on the chin and sides of the cheek, and some information might be at the upper cheeks and the sides of the nose (Preminger et al., 1998). Campbell (2001) applied an independent component analysis to speaker video data and showed that lips and teeth are functional cues for lip-reading, but so are jaw, skin wrinkling, neck, and chin movement. The importance of these cues was then validated in a perception experiment. It should not be surprising that facial areas outside the mouth region are informative about the speech signal, since the muscles connected to the lips also consequently move other regions. For example, Munhall and Vatikiotis-Bateson (1998; see also Yehia et al., 1998) investigated the correlation of movements of different regions of the face. They found almost perfect positive correlation ($r \geq .95$) between the movement of the mouth and other regions in and outside the face (e.g. cheeks). However, areas outside the mouth contain information independent from the mouth (Vatikiotis-Bateson and Yehia, 1996; reported in Munhall and Vatikiotis-Bateson, 2004). This is in agreement with studies showing that only about 40–60 percent of all eye fixations during visual and audiovisual speech perceptions are on the mouth region (Vatikiotis-Bateson et al., 1998). How often a listener fixates on the mouth region seems to depend on the quality of the auditory signal (i.e. increasing mouth fixations with decreasing signal-to-noise ratio; Vatikiotis-Bateson et al., 1998) and the task (i.e. more fixations on the eye level for prosodic categorization than for segmental identification tasks; Lansing and McConkie, 1999). However, fixation at a certain region cannot be interpreted as a measure of where the information in the face is located. Even parafoveal vision provides sufficient visual information, for example, to produce McGurk responses. Fixations on regions outside the mouth do not impact the influence of visual speech information on incongruent audiovisual speech perception (Paré et al., 2003).

But what information is transmitted by these regions other than the mouth? Smeele (1994) examined the importance of jaw, lips, and oral cavity in producing a McGurk effect. She showed that the presentation of either the oral cavity or the lips is sufficient to obtain McGurk combination responses, but almost no fusion responses. In order to obtain fusion responses, both lip and oral cavity information was necessary. McGurk illusions rarely occurred if only information from the jaw was presented with the auditory signal. It seems that jaw movement alone does not provide sufficient place of articulation information in order for visual speech to have a noticeable impact on the overall percept. On the other hand, the movement of other non-oral areas of the face seems to carry enough information to impact audiovisual speech perception (Preminger et al., 1998; Thomas and Jordan, 2004). For example, when participants were presented with a video-manipulated speaker who only moved the areas outside the mouth but not the mouth region itself, a substantial influence of visual speech information on audiovisual speech recognition (as measured by McGurk responses) was found (Thomas and Jordan, 2004). It seems therefore that, while the addition of seeing the jaw of a speaker is not sufficient to contribute place of articulation information, the movement of the head as a whole is. However, information from the mouth region alone transmits usually sufficient visual information. Seeing the face moving in addition to the mouth does not add to the influence of visual speech on the audiovisual percept, nor does it further improve the recognition of visual-only speech (Thomas and Jordan, 2004; see also Ijsseldijk, 1992; Marassa and Lansing, 1995). Furthermore, seeing a static mouth in combination with the moving face precludes the influence of visual speech on audiovisual speech perception compared to seeing a moving face alone (Thomas and Jordan, 2004).

Two basic questions related to this research need to be addressed. One question is whether visual speech contributes to perception through featural cues (e.g. through the correlation between movements of the lips and other facial regions) or through configurational cues (i.e. the relative spatial distance between two or more features in the face, e.g. the chin and the nose during articulation). The second question is whether static information (i.e. fixed configurations of the face during speech) or dynamic information plays a role in speech reading and audiovisual speech perception. Degrading featural information, for example by blurring the video image, decreases accuracy in visual speech perception as well as the influence of visual speech on the perception of congruent and incongruent audiovisual speech in noise and in quiet (Thomas and Jordan, 2002). Similarly, degrading configurational information by rotating or completely inverting the whole face reduces the influence of visual speech information (Jordan and Bevan, 1997; Massaro and Cohen, 1996; Rosenblum et al., 2000; but see Thomas and Jordan, 2002). However, the influence of inversion on speech perception might be stimulus-dependent (Massaro and Cohen, 1996; Rosenblum et al., 2000). Future research is necessary to investigate this issue more thoroughly, but it appears that the configurational change through rotation only impacts the information transmitted, not the processing of information (Massaro and Cohen, 1996). Overall, it is also to be noted that while speech perception seems to be impacted by changes in featural and configurational information, such as through blurring and rotation, audiovisual speech perception is generally quite robust against the impact of these variables.

With regard to the second basic question—whether seeing the speaker contributes static or dynamic information to visual and audiovisual speech perception—it can be said that while static information can be sufficient for speech recognition (Campbell, 1986; Campbell et al., 1986), the dynamics of the face and lips are also informative about the visual percept. Brooke and Summerfield (1983) showed that vowels are produced with characteristic lip and jaw trajectories. These trajectories are distinguishable on account of their difference in movement, velocity, and acceleration. Analogous to classic studies of motion perception (Johansson, 1973), Rosenblum et al. (1996) tested the influence of time-varying information by replacing the face with point-lights that were positioned strategically on the face. The configurations that were tested had point-lights only at the lips, or at the lips and at the teeth and tongue. A third configuration had point-lights at lips, teeth, tongue, and on the face. An audiovisual recognition benefit was found for the presentation with lights only on the lips. However, performance in presentation conditions with lights on lips, teeth, and tongue was substantially better than in conditions with lights only on the lips. The points on the face did not add to the audiovisual benefit above and beyond the benefit found for the condition with point-lights on lips, teeth, and tongue (Rosenblum et al., 1996). Time-varying information was also sufficient to produce the McGurk illusion (Rosenblum and Saldaña, 1996). However, performance with point light displays is significantly poorer than it is with the full face displays (Cohen et al., 1996).

In summary, audiovisual speech perception is quite robust vis-à-vis different forms of image distortions (e.g. blurring, viewing angle, or inversion). This might be due to the fact that although the lips are the main contributors to visual information, other non-oral areas of the face also transmit information. This information is partially redundant with the information from the lips, but also partially unique. Furthermore, the type of information provided by visual speech can be robust against visual distortions.

2.3 Theories of audiovisual speech perception

A theory of audiovisual speech perception needs to describe the visual psychophysics of visual speech as well as how visual speech is evaluated and integrated with auditory speech. Furthermore, it needs to explain to what extent this additional source of information is considered when making a decision about what was perceived. It thereby needs to explain how these processes account for the audiovisual benefit, the McGurk effect, and the other phenomena described in the previous section. The theory can either postulate processes that are unique to speech recognition or try to explain audiovisual speech recognition as part of a general approach to pattern recognition. Although most of the influential theories of auditory spoken word recognition (see Gaskell, Chapter 4 this volume), such as TRACE (McClelland and Elman, 1986), Cohort theory (Marslen-Wilson, 1987), Merge (Norris et al., 2000), or the Neighborhood Activation Model (Luce and Pisoni, 1998), do not deal with visible speech, this neglect seems to be a restriction in research focus rather than reflecting the assumption that visual speech has no influence in face-to-face communication.

2.3.1 Psychoacoustic accounts

Psychoacoustic theories discount the influence of visual speech. This class of theories assumes that speech processing is nothing else than the processing of complex sounds. More recent versions of psychoacoustic theories now acknowledge the influence of visual speech (e.g. Diehl and Kluender, 1987), but lack an account of how and to what degree visible speech is influential. Not surprisingly, psychoacoustic theorists could only give a secondary role to visual speech in order to preserve the very nature of psychoacoustic theories: that acoustics are central in understanding spoken language. As we have seen in the examples described above, however, visible speech cannot be relegated to a secondary role in a satisfactory account of audiovisual speech perception.

2.3.2 Motor theory

The basic assumption of the motor theory is that speech is perceived in terms of gestures (Liberman and Mattingly, 1985; Mattingly and Studdert-Kennedy, 1991). The perceiver attempts to recover the articulatory gestures that produced the speech in order to understand what was said. This assumption was motivated by solving the "invariance problem" that there is no identifiable one-to-one correspondence between the acoustic signal and phonemes that could contribute to speech recognition. It is argued by the motor theory, however, that there is a lawful relationship between specific phonemes and the articulatory gestures used to produce them. Supposedly, due to coarticulation with the subsequent vowel, the same consonant can be acoustically different, but will be articulatorily somewhat similar.

As a consequence of this gestural mediation assumption, speech perception is postulated by motor theorists to be special and different from auditory perception in its processes and information sources (Liberman, 1996). This phonetic module is specific to auditory as well as visual speech. Both sources are integrated through the common gestural representation of both visual and auditory speech signals. However, research has shown that the same processes involved in other pattern recognition domains can also account for the integration of multiple sources of information in speech perception (Massaro, 1998). There is no need to postulate a special processing module for speech. There is little empirical evidence that directly support the mediation of speech perception by gestures.

2.3.3 Direct perception

Direct perception theory also claims that gestures are the primary objects of speech perception, but does not assume that speech is special (Fowler 1986; 1996). Rather, the direct perception theory for speech perception is placed within Gibson's (1966) framework of direct perception, which postulates that persons directly perceive the causes of sensory input. In spoken language, the cause of an audible-visible speech percept is the vocal-tract activity of the talker. Accordingly, it is reasoned that visible speech should influence speech perception because it also reveals the vocal-tract activity of the talker. Speech perceivers therefore obtain direct information from integrated perceptual systems from the flow of stimulation provided by the talker (Best, 1995). The observed influence of visible speech is easily predicted by this theory because visible speech represents another source of stimulation, providing direct information about the gestural actions of the talker. However, we know of no convincing evidence for the gesture as the primary object of speech perception (see Massaro, 1998; Ohala, 1996). For now, it seems most parsimonious to assume that the objects of speech perception are relatively abstract symbols (Nearey, 1992; but see Pisoni and Levi, Chapter 1 this volume, for a discussion).

2.3.4 Pattern recognition accounts

Speech perception can also be described as a case of pattern recognition that involves the evaluation of all available sources of information as well as the integration of this information in order to reach a decision on what was perceived. The underlying processes of pattern recognition would be domain-general; however, the sources of information involved in recognition would most likely be domain-specific. Visual speech is therefore simply another source of information that is considered to understand what was said. The information provided from watching the speaker will be integrated with all other available information. Integration is a general algorithm that applies to all available sources of information. In speech, these other sources could involve, among others, lexical or context knowledge, or knowledge about phonology, syntax, semantics, or pragmatics. The same algorithm operates on the integration of multiple cues independently of whether these cues were all obtained from the same modality or from different modalities.

The Fuzzy Logical Model of Perception (FLMP, Massaro, 1998) gives such a pattern recognition account of speech perception. According to the FLMP, all sources of information are evaluated independently of each other and integrated. The output of the evaluation process is knowledge about the degree to which each source of information matches the prototypes for each possible alternative as stored in long-term memory. Prototypes are summary descriptions of the best exemplars of a category. Fuzzy truth values are the common metric in which the degree of support from all sources is expressed. Support values are expressed as fuzzy truth values ranging between zero and one. Values at the .5 level represent complete ambiguity. In other words, if one were to create a stimulus that is completely ambiguous on a continuum between /ba/ and /da/, then the support value would be .5. With an increase or decrease of this value away from .5, ambiguity decreases. Integration of information follows a multiplicative combination of these support values. When combined in a relative decision rule, this allows accounting for better performance based on two imperfect sources than on considering only either one. For example, given a two-alternative forced-choice task between /da/ and /ba/, the degree of support provided by the auditory information for /da/ can be noted as a_i and the support for /ba/ as $(1- a_i)$. Similarly, the support provided by the visual information for /da/ is written as v_j, and the support for /ba/ as $(1-v_j)$. The overall support for /da/ from visual and auditory speech information is the product of the support value v_j and a_i. All available information is used to form a decision. There is no information loss due to integration.

Perceptual identification is, however, based on the degree of overall support for one alternative relative to the summed overall support for all other alternatives. For example, the relative degree of support for /da/ is equal to the overall support for /da/ divided by the sum of the overall support for all alternatives, namely here for /da/ and /ba/.

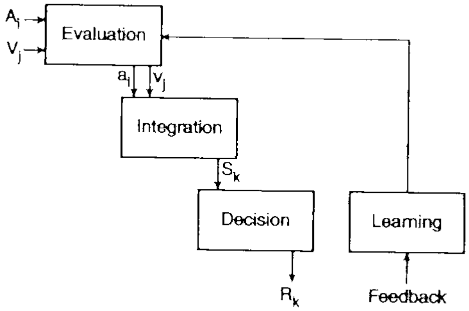

Figure 2.1 Schematic representation of the three processes involved in perceptual recognition. The three processes are shown to precede left to right in time to illustrate their necessarily successive but overlapping processing. These processes make use of prototypes stored in long-term memory. The sources of information are represented by uppercase letters. Auditory information is represented by A_i and visual information by V_j. The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters a_i and v_j) These sources are then integrated to give an overall degree of support, S_k, for each speech alternative k. The decision operation maps the outputs of integration into some response alternative, R_k. The response can take the form of a discrete decision or a rating of the degree to which the alternative is likely. The feedback is assumed to tune the prototypical values of the features used by the evaluation process.

$$P(/da/\,|\,A_iV_j)=\frac{a_iv_j}{a_iv_j+(1-a_i)(1-v_j)} \qquad (1)$$

The FLMP predicts that the influence of a source of information on the decision increases to the degree that other sources are ambiguous. Cognitive sources of information (e.g. context knowledge) are treated exactly the same way by this information-processing model. Their independently evaluated support for each alternative is combined multiplicatively with the degree of support from all other sources. Therefore, the FLMP does not assume that cognitive sources of information modify the perceptual input, but rather that all information is solely passed on and integrated to determine the overall relative support for each alternative. This central assumption stands in marked contrast to interactive activation models and anticipated the Merge model (Norris et al., 2000; see Massaro, 2000 for a commentary).

Given the quantitative nature of the FLMP, it provides better testing conditions than verbal models like the motor theory. The FLMP has been broadly evaluated in different areas, tasks, materials, and participant populations (see Massaro, 1998 for an overview). This provides evidence that although the amount and type of information between studies differs, the processing of this information adheres to the same principles of pattern recognition as outlined by the FLMP. The integration algorithm of the FLMP has been evaluated by comparing the ability of the model to fit behavioral data against model implementations with alternative integration rules (see Massaro, 1998 for an overview). For example, in contrast to the FLMP, an averaging integration model would predict that the overall degree of support can never exceed the degree of support provided by the most informative source of information. Therefore, audiovisual speech can never be more informative than unimodal speech. An additive model (see e.g. Cutting et al., 1992; Cutting and Bruno, 1988) would make the same predictions as an averaging model when combined with the relative goodness rule for decision. The FLMP provides a better account of pattern recognition than competitor models that had an averaging or an additive rule of integration (e.g. Massaro and Cohen, 1993b; Massaro, 1998).

Given this success of the FLMP's integration algorithm, a natural question is: what is the postulated underlying neural mechanism? Generally, there are at least three different neurologically plausible solutions for how integration of visible speech with auditory speech might occur. Auditory and visual neurons transmit modality-specific information to other neurons. The three solutions differ with respect to how the auditory and visual information is shared during the chain of processing from input to output. In the first case, the processing of the visual modality activates the location that receives activation from the other modality. We call this "sensory penetration" because information from the visual modality impacts on the neurons processing the auditory modality. As illustrated in Figure 2.2a, the activation from the visible speech is sent to a location whose primary function is to receive activation from the auditory modality. This neurological instantiation represents a so-called non-independence model of integration in which one modality is never represented independently of the other modality.

A second type of brain integration involves simple feed-forward convergence. As illustrated in Figure 2.2b, the neural activation from the auditory and visible speech activates a third location that is sensitive to the inputs from both modalities. An important set of observations from single-cell recordings in cats could be interpreted in terms of convergent integration (Stein and Meredith, 1992).

Non-convergent temporal integration, illustrated in Figure 2.2c, involves integration-like behavior, but there is no location at which the separate modality-specific information is integrated. This type of integration involves the combination of information from two or more

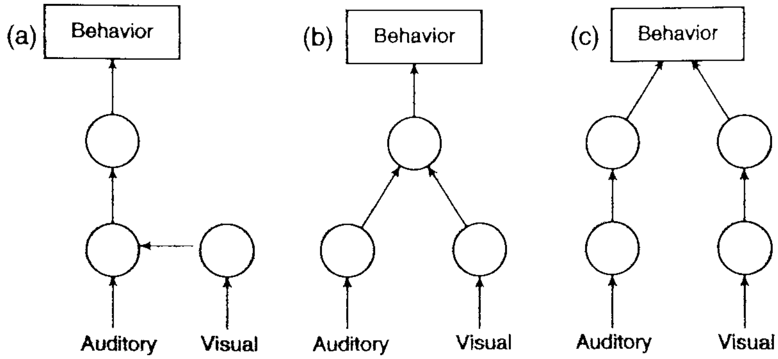

Figure 2.2 Schematic representation of the neural processing involved in sensory penetration (2a), simple feedforward convergence (2b), and nonconvergent temporal integration (2c).

remote regions of the brain. Corticocortical pathways (pathways that connect regions of the cortex) synchronize the outputs of these regions and enable them to feed forward, independently but synchronously, to other areas. This type of integration would influence some output process rather than producing some higher-order integrated representation for storage or further processing by other parts of the brain.

The FLMP is consistent with the latter two models, but not the first one, by predicting the simultaneous but independent influence of both audible and visible speech. What is sufficient for an implementation of the FLMP is neural activity along two independent channels that simultaneously influence behavior.

2.4 Visual speech as a language learning tool

As the previous sections outlined, visual speech contributes to the robust and successful recognition of speech. Therefore, it holds promise for educational applications such as language tutoring for second-language learners, for children and adults with hearing impairment, and for autistic children. Language tutoring that includes presentations of visual speech is not only critical for speech-reading acquisition, and the production training of new and old vocabulary, but also for the general acquisition of new vocabulary (Bosseler and Massaro, 2003; Hardison, 2003).

It has been shown that speech-reading can be improved by training (Massaro et al., 1993b; Walden et al., 1977). When trained to recognize consonant-vowel syllables presented auditory-only, visual-only, or audiovisually, participants improved in all three presentation conditions (Massaro et al., 1993b). Results of two recent experiments show that participants seem to learn the same amount during visual-only presentations with auditory speech feedback and audiovisual speech-reading training (Geraci and Massaro, 2002). From the research presented throughout this chapter, we can conclude that humans naturally can speech-read, but also can improve when trained. It follows that speech-reading is a valuable skill for robust speech recognition and can be enhanced when needed, for example, to compensate for hearing difficulties. When trained with supplementary visual speech, hard of hearing children between the ages of 8 and 13 improved their speech perception and production skills (Massaro and Light, 2004). Their acquired production skills generalized to a new set of words that had not been included in the training sessions. The speech production skills deteriorated somewhat after six weeks without training. These results are evidence that the training method, rather than some other experiences, was responsible for the learning.

Multimodal speech training has also been shown to be effective for vocabulary acquisition of children with autism. This is despite the evidence showing that autistic children are impaired in face processing (Dawson et al., 2002; Rogers 1999; Williams et al., 2001) and tend to avoid face-to-face contact with others (Happe, 1998). However, de Gelder and colleagues (1996) reported evidence that children with autism are less influenced by visual speech in audiovisual speech recognition paradigms than normally developing children matched for mental age. But the smaller visual effect for children with autism does not necessarily reflect non-optimal integration, but rather could be due to less information from the face. The latter explanation is supported by reanalysis of the data of de Gelder et al. within the framework of the FLMP (Massaro and Bosseler, 2003), which showed that the children with autism seem to apply the same optimal integration algorithm as their comparison group. Furthermore, the behavioral data showed that audiovisual speech training can be effective in teaching new vocabulary and speech-reading to autistic children (Massaro and Bosseler, 2003). A more recent investigation provided evidence that indeed the face facilitated this learning (Massaro and Bosseler, 2006).

Information provided by the face is also helpful for the acquisition of a foreign language. Second-language learners would benefit from audiovisual presentations not only for the purpose of learning new vocabulary but also for accent reduction training. For example, native Japanese speakers benefited from facial information when learning the perception and production of the American English /r/ and /l/ (Massaro and Light, 2003). The successful training also led to generalization of the skill to new words.

2.5 Conclusion

Although we often talk at a hidden distance, the absence of the face reduces information support for understanding. Multimodal conversation in face-to-face situations reveals the complexity and richness of language and the challenge of uncovering the processes that allow us to communicate so seamlessly. This chapter is unique in this handbook because it specifically addresses the multimodal nature of perception (see also

Frost and Ziegler, Chapter 7 this volume) and word recognition in speech. We encourage the reader to apply some of the concepts developed in this chapter in their reading of other related chapters. What challenges do other approaches provide, and how might they be tested within our framework? Analogously, are there concepts and findings that can enrich the other descriptions of language processing found within this volume?

References

Arnold, P., and Hill, F. (2001) Bisensory augmentation: a speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, 92: 339–55.

Auer, E. T. Jr (2002) The influence of the lexicon on speech read word recognition: contrasting segmental and lexical distinctiveness. *Psychonomic Bulletin and Review*, 9: 341–7.

Auer, E. T. Jr, and Bernstein, L. E. (1997) Speechreading and the structure of the lexicon: computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. *Journal of the Acoustical Society of America*, 102: 3704–10.

Benguerel, A. P., and Pichora-Fuller, M. K. (1982) Coarticulation effects in lipreading. *Journal of Speech and Hearing Research*, 25: 600–607.

Benoît, C., Guiard-Marigny, T., Le Goff, B., and Adjoudani, A. (1996) Which components of the face do humans and machines best speechread? In D. G. Stork and M. E. Hennecke (eds), *Speechreading by Humans and Machines: Models, Systems, and Applications*, pp. 315–25. Springer, Berlin.

Besle, J., Fort, A., Delpuech, C., and Giard, M.-H. (2004) Bimodal speech: early suppressive visual effects in human auditory cortex. *European Journal of Neuroscience*, 20: 2225–34.

Best, C. T. (1995) A direct realist perspective on cross-language speech perception. In W. Strange (ed.), *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research*, pp. 167–200. York Press, Timonium, MD.

Bosseler, A., and Massaro, D. W. (2003) Development and evaluation of a computer- animated tutor for vocabulary and language learning for children with autism. *Journal of Autism and Developmental Disorders*, 33: 654–73.

Braida, L. D. (1991) Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*. Special Issue: *Hearing and Speech*, 43: 647–77.

Brancazio, L. (1999) Contributions of the lexicon to audiovisual speech perception. *Dissertation Abstracts International*, 59: 5591B. (UMI No. 9909779).

Brancazio, L. (2004) Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 30: 445–63.

Breeuwer, M., and Plomp, R. (1986) Speechreading supplemented with auditorily presented speech parameters. *Journal of the Acoustical Society of America*, 79: 481–99.

Brooke, N. M., and Summerfield, A. Q. (1983) Analysis, synthesis and perception of visible articulatory movements. *Journal of Phonetics*, 11: 63–76.

Campbell, C. S. (2001) Patterns of evidence: investigating information in visible speech perception. *Dissertation Abstracts International*, 61: 3869B (UMI No. 9979917).

Campbell, C. S., and Massaro, D. W. (1997) Perception of visible speech: Influence of spatial quantization. *Perception*, 26: 627–44.

Campbell, R. (1986) The lateralization of lipread sounds: a first look. *Brain and Cognition*, 5: 1-21.

Campbell, R. and Dodd, B. (1980) Hearing by eye. *Quarterly Journal of Experimental Psychology*, 32: 85–99.

Campbell, R., Landis, R., and Regard, M. (1986) Face recognition and lipreading: a neurological dissociation. *Brain*, 109: 509–21.

Cathiard, M.-A., Lallouache, M. T., Mohamadi, T., and Abry, C. (1995) Configurational vs. temporal coherence in audio-visual speech perception. *Proceedings of the 13th International Congress of Phonetic Sciences*, 3: 218–21.

Cohen, M. M., Walker, R. L., and Massaro, D. W. (1996) Perception of synthetic visual speech. In D. G. Stork and M. E. Hennecke (eds), *Speechreading by Humans and Machines: Models, Systems, and Applications*, pp. 153–68. Springer, Berlin.

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., and Deltenre, P. (2002) Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. *Clinical Neurophysiology*, 113: 495–506.

Connine, C. M., Blasko, D. G., and Titone, D. (1993) Do the beginnings of spoken words have a special status in auditory word recognition? *Journal of Memory and Language*, 32: 193–210.

Cutting, J. E., and Bruno, N. (1988) Additivity, subadditivity, and the use of visual information: a reply to Massaro (1988). *Journal of Experimental Psychology: General*, 117: 422–4.

Cutting, J. E., Bruno, N., Brady, N. P., and Moore, C. (1992) Selectivity, scope, and simplicity of models: a lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, 121: 364–81.

Dawson, G., Carver, L., Meltzoff, A. N., Panagiotides, H., McPartland, J., and Webb, S. J. (2002) Neural correlates of face and object recognition in young children with autism spectrum disorder, developmental delay, and typical development. *Child Development*, 73: 700–717.

de Gelder, B., and Vroomen, J. (2000) The perception of emotion by ear and by eye. *Cognition and Emotion*, 14: 289–311.

de Gelder, B., Vroomen, J., and van der Heide, L. (1996) Face recognition and lip-reading in autism. *European Journal of Cognitive Psychology*, 3: 69–86.

de la Vaux, S. K., and Massaro, D. W. (2004) Audiovisual speech gating: examining information and information processing. *Cognitive Processing*, 5: 106–12.

Diehl, R. L., and Kluender, K. R. (1987) On the categorization of speech sounds. In S. Harnad (ed.), *Categorical perception*, pp. 226–53. Cambridge University Press, Cambridge.

Ellison, J. W., and Massaro, D. W. (1997) Featural evaluation, integration, and judgment of facial affect. *Journal of Experimental Psychology: Human Perception and Performance*, 23: 213–26.

Erber, N. P. (1971) Effects of distance on the visual reception of speech. *Journal of Speech and Hearing Research*, 14: 848–57.

Erber, N. P. (1974) Effects of angle, distance, and illumination on visual reception of speech by profoundly deaf children. *Journal of Speech and Hearing Research*, 17: 99–112.

Fisher, C. G. (1968) Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 15: 474–82.

Fowler, C. A. (1986) An event approach to the study of speech perception from a direct realist perspective. *Journal of Phonetics*, 14: 3–28.

Fowler, C. A. (1996) Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, 99: 1730–41.

Ganong, W. F. (1980) Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, 6: 110–25.

Geraci, K., and Massaro, D. W. (2002) Teaching speechreading: is unimodal or bimodal training more effective? MS.

Gibson, J. J. (1966) *The Senses Considered as Perceptual Systems*. Houghton Mifflin, Boston.

Grant, K. W., and Greenberg, S. (2001) Speech intelligibility derived from asynchronous processing of auditory-visual information. In D. W. Massaro, J. Light, and K. Geraci (eds), *Proceedings of the AVSP 2001*, pp. 132–7. Aalborg, Denmark.

Grant, K. W., and Seitz, P. F. (2000) The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, 108: 1197–1208.

Grant, K. W., Walden, B. E., and Seitz, P. F. (1998) Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America*, 103: 2677–90.

Green, K. P., and Gerdman, A. (1995) Cross-modal discrepancies in coarticulation and the integration of speech information: the McGurk effect with mismatched vowels. *Journal of Experimental Psychology: Human Perception and Performance*, 21: 1409–26.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., and Stevens, E. B. (1991) Integrating speech information across talkers, gender and sensory modality: female faces and male voices in the McGurk effect. *Perception and Psychophysics*, 50: 524–36.

Greenberg, S. (2005) A multi-tier framework for understanding spoken language. In S. Greenberg and W. Ainsworth (eds), *Listening to Speech: An Auditory Perspective*. Erlbaum, Hillsdale, NJ.

Greenberg, S., and Arai, T. (2004) What are the essential cues for understanding spoken language? *IEICE Transactions on Information and Systems*, E87-D(5): 1059–70.

Happe, F. (1998) *Autism: An Introduction to Psychological Theory*. Harvard University Press, Cambridge, MA.

Hardison, D. M. (2003) Acquisition of second-language speech: effects of visual cues, context, and talker variability. *Applied Psycholinguistics*, 2: 495–522.

Ijsseldijk, F. J. (1992) Speechreading performance under different conditions of video image, repetition, and speech rate. *Journal of Speech and Hearing Research*, 35: 46–71.

Jackson, P., Montgomery, A. A., and Binnie, C. A. (1976) Perceptual dimensions underlying vowel lipreading performance. *Journal of Speech and Hearing*, 19: 796–812.

Jesse, A. (2005) Towards a lexical fuzzy logical model of perception: the time-course of information in lexical identification of face-to-face speech. Doctoral dissertation, University of California, Santa Cruz.

Jesse, A., Vrignaud, N., Cohen, M. M., and Massaro, D. W. (2000/2001) The processing of information from multiple sources in simultaneous interpreting. *Interpreting*, 5: 95–115.

Johansson, G. (1973) Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14: 201–11.

Jordan, T. R., and Bevan, K. M. (1997) Seeing and hearing rotated faces: influences of facial orientation on visual and audiovisual speech recognition. *Journal of Experimental Psychology: Human Perception and Performance*, 23: 388–403.

Jordan, T., and Sergeant, P. (2000) Effects of distance on visual and audiovisual speech recognition. *Language and Speech*, 43: 107–24.

Kewley-Port, D. (1983) Time varying features as correlates of place of articulation in stop consonants. *Journal of the Acoustical Society of America*, 73: 322–35.

Lansing, C. R., and McConkie, G. W. (1999) Attention to facial regions in the segmental and prosodic visual speech perception tasks. *Journal of Speech, Language, and Hearing Research*, 42: 526–39.

Lewald, J., and Guski, R. (2003) Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. *Cognitive Brain Research*, 16: 468–78.

Liberman, A. M. (1996) *Speech: A Special Code*. MIT Press, Cambridge, MA.

Liberman, A. M., and Mattingly, I. (1985) *The* motor theory of speech perception revised. *Cognition*, 21: 1–36.

Luce, P. A., and Pisoni, D. B. (1998) Recognizing spoken words: the neighborhood activation model. *Ear and Hearing*, 19: 1–36.

MacDonald, J., Andersen, S., and Bachmann, T. (2000) Hearing by eye: how much spatial degradation can be tolerated? *Perception*, 29: 1155–68.

MacDonald, J., and McGurk, H. (1978) Visual influences on speech perception processes. *Perception and Psychophysics*, 24: 253–7.

MacLeod, A., and Summerfield, Q.(1987) Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology,* 21: 131–41.

Marassa, L. K., and Lansing, C. R. (1995) Visual word recognition in 2 facial motion conditions: full face versus lips-plus-mandible. *Journal of Speech and Hearing Research,* 38: 1387–94.

Marslen-Wilson, W. D. (1987) Functional parallelism in spoken word-recognition. *Cognition,* Special Issue: *Spoken Word Recognition,* 25: 71–102.

Marslen-Wilson, W. D., Moss, H. E., and van Halen, S. (1996) Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance,* 22: 1376–92.

Marslen-Wilson, W., and Zwitserlood, P. (1989) Accessing spoken words: the importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance,* 15: 576–85.

Massaro, D. W. (1987) *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry.* Erlbaum, Hillsdale, NJ.

Massaro, D. W. (1998) *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle.* MIT Press, Cambridge, MA.

Massaro, D. W. (2000) The horse race to language understanding: FLMP was first out of the gate, and has yet to be overtaken. *Behavioral and Brain Sciences,* 23: 338–9.

Massaro, D. W. (2003) Model selection in AVSP: some old and not so old news. In J. L. Schwartz, F. Berthommier, M. A. Cathiard, and D. Sodoyer (eds), *Proceedings of Auditory-Visual Speech Processing Conference,* pp. 83–8. St Jorioz, France.

Massaro, D. W., and Bosseler, A. (2003) Perceiving speech by ear and eye: multimodal integration by children with autism. *Journal of Developmental and Learning Disorders,* 7: 111–44.

Massaro, D. W., and Bosseler, A. (2006) Read my lips: the importance of the face in a computer-animated tutor for autistic children learning language. *Autism: The International Journal of Research and Practice.*

Massaro, D. W., and Cohen, M. M. (1993a) Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication,* 13: 127–34.

Massaro, D. W., and Cohen, M. M. (1993b) The *paradigm* and the fuzzy logical model of perception are alive and well. *Journal of Experimental Psychology: General,* 122: 115–24.

Massaro, D. W., and Cohen, M. M. (1996) Perceiving speech from inverted faces. *Perception and Psychophysics,* 58: 1047–65.

Massaro, D. W., and Cohen, M. M. (1999) Speech perception in hearing-impaired perceivers: synergy of multiple modalities. *Journal of Speech, Language, and Hearing Science,* 42: 21–41.

Massaro, D. W., Cohen, M. M., and Gesi, A. T. (1993b). Long-term training, transfer, and retention in learning to lipread. *Perception and Psychophysics,* 53: 549–62.

Massaro, D. W., Cohen, M. M., and Smeele, P. M. T. (1996) Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America,* 100: 1777–86.

Massaro, D. W., Cohen, M. M., Gesi, A. T., and Heredia, R. (1993a) Bimodal speech perception: an examination across languages. *Journal of Phonetics,* 21: 445–78.

Massaro, D. W., and Egan, P. B. (1996) Perceiving affect from the voice and the face. *Psychonomic Bulletin and Review,* 3: 215–21.

Massaro, D. W., and Light, J. (2003) Read my tongue movements: bimodal learning to perceive and produce non-native speech /r/ and /l/. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech'03/ Interspeech'03)*(CD-ROM, 4 pp.). Geneva.

Massaro, D. W., and Light, J. (2004) Using visible speech for training perception and production of speech for hard of hearing individuals. *Journal of Speech, Language, and Hearing Research,* 47: 304–20.

Mattingly, I. G., and Studdert-Kennedy, M. (eds) (1991) *Modularity and the Motor Theory of Speech Perception.* Erlbaum, Hillsdale, NJ.

Mattys, S. L., Bernstein, L. E., and Auer, E. T. Jr (2002) Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Perception and Psychophysics,* 64, 667–79.

McClelland, J. L., and Elman, J. L. (1986) The TRACE model of speech perception. *Cognitive Psychology,* 18: 1–86.

McGurk, H., and MacDonald, J. (1976) Hearing lips and seeing voices. *Nature,* 264: 746–8.

Miller, G. A., and Nicely, P. A. (1955) An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America,* 27: 338–52.

Mills, A. E., and Thiem, R. (1980) Auditory-visual fusions and illusions in speech perception. *Linguistische Berichte,* 68/80: 85–108.

Montgomery, A. A., and Jackson, P. L. (1983) Physical characteristics of the lips underlying vowel lipreading performance. *Journal of the Acoustical Society of America,* 73: 2134–44.

Munhall, K. G., Gribble, P., Sacco, L., and Ward, M. (1996) Temporal constraints on the McGurk effect. *Perception and Psychophysics,* 58: 351–62.

Munhall, K. G., Kroos, C., Jozan, G., and Vatikiotis-Bateson, E. (2004) Spatial frequency requirements for audiovisual speech perception. *Perception and Psychophysics,* 66: 574–83.

Munhall, K. G., and Tohkura, Y. (1998) Audiovisual gating and the time course of speech perception. *Journal of the Acoustical Society of America,* 104: 530–9.

Munhall, K. G., and Vatikiotis-Bateson, E. (1998) The moving face during speech communication. In B. Dodd, R. Campbell, and D. Burnham (eds), *Hearing by Eye, part 2: The Psychology of Speechreading and Audiovisual Speech,* pp. 123–39. Taylor & Francis, London.

Munhall, K. G., and Vatikiotis-Bateson, E. (2004) Spatial and temporal constraints on audiovisual speech perception. In G. A. Calvert, C. Spence, and B. E. Stein (eds), *The Handbook of Multisensory Processes,* pp. 117–88. MIT Press, Cambridge, MA.

Nearey, T. M. (1992) Context effects in a double-weak theory of speech perception. *Language and Speech*, 35: 153–71.

Norris, D., McQueen, J. M., and Cutler, A. (2000) Merging information in speech recognition: feedback is never necessary. *Behavioral and Brain Sciences*, 23: 299–370.

Ohala, J. J. (1996) Speech perception is hearing sounds, not tongues. *Journal of the Acoustical Society of America*, 99: 1718–25.

Ouni, S., Cohen, M. M., Ishak, H., and Massaro, D. W. (2005) Visual contribution to speech perception: measuring the intelligibility of talking heads. *Proceedings of the Auditory-Visual Speech Processing Conference*, pp. 45–46. British Columbia, Canada.

Paré, M., Richler, R., ten Hove, M., and Munhall, K. G. (2003) Gaze behavior in audiovisual speech perception: the influence of ocular fixations on the McGurk effect. *Perception and Psychophysics*, 65: 553–67.

Preminger, J. E., Lin, H. B., Payen, M., and Levitt, H. (1998) Selective visual masking in speechreading. *Journal of Speech, Language and Hearing Research*, 41: 564–75.

Reisberg, D., McLean, J., and Goldfield, A. (1987) Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli. In B. Dodd and R. Campbell (eds), *Hearing by Eye: The Psychology of Lip-Reading*, pp. 97–113. Erlbaum, Hillsdale, NJ.

Rogers, S. J. (1999) Intervention for young children with autism: from research to practice. *Infants and Young Children*, 12: 1–16.

Rosenblum, L. D., Johnson, J. A., and Saldaña, H. M. (1996) Visual kinematic information for embellishing speech in noise. *Journal of Speech and Hearing Research*, 39: 1159–70.

Rosenblum, L. D., and Saldaña, H. M. (1996) An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 22: 318–31.

Rosenblum, L. D., Schmuckler, M. A., and Johnson, J. A. (1997) The McGurk effect in infants. *Perception and Psychophysics*, 59: 347–57.

Rosenblum, L. D., Yakel, D. A., and Green, K. P. (2000) Face and mouth inversion effects on visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 26: 806–19.

Sams, M., Mötönen, R., and Sihvonen, T. (2005) Seeing and hearing others and oneself talk. *Cognitive Brain Research*, 23: 429–35.

Sams, M., Surakka, V., Helin, P., and Kättö, R. (1997) Audiovisual fusion in Finnish syllables and words. *Proceedings of the Auditory-Visual Speech Processing Conference*, pp. 101–4. Rhodes, Greece.

Schwartz, J.-L. (2003) Why the FLMP should not be applied to McGurk data … or how to better compare models in the Bayesian framework. *Proceedings of the Audiovisual Speech Perception Conference*, pp. 77–82. St Jorioz, France.

Schwartz, J.-L., Berthommier, F., and Savariaux, C. (2004) Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, 93: B69–B78.

Sekiyama, K., and Tohkura, Y. (1991) McGurk effect in non-English listeners: few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America*, 90: 1797–1805.

Sekiyama, K., and Tohkura, Y. (1993) Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, 21: 427–44.

Shannon, C. E. (1948) A mathematical theory of communications. *Bell Systems Technical Journal*, 27: 379–423.

Shillcock, R. (1990) Lexical hypotheses in continuous speech. In G. T. M. Altmann (ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*, pp. 24–49. MIT Press, Cambridge, MA.

Smeele, P. M. T. (1994) Perceiving speech: integrating auditory and visual speech. Doctoral dissertation, Delft University of Technology.

Stein, B. E., and Meredith, M. A. (1992) *The Merging of the Senses*. MIT Press, Cambridge, MA.

Sumby, W. H., and Pollack, I. (1954) Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26: 212–15.

Summerfield, A. Q. (1979) Use of visual information in phonetic perception. *Phonetica*, 36: 314–31.

Summerfield, A. Q. (1987) Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd and R. Campbell (eds), *Hearing by Eye: The Psychology of Lip-Reading*, pp. 3–51. Erlbaum, London.

Summerfield, A. Q., and McGrath, M. (1984) Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology*, 36A: 51–74.

Tabossi, P., Burani, C., and Scott, D. (1995) Word identification in fluent speech. *Journal of Memory and Language*, 34: 440–67.

Thomas, S. M., and Jordan, T. R. (2002) Determining the influence of Gaussian blurring on inversion effects with talking faces. *Perception and Psychophysics*, 64: 932–44.

Thomas, S. M., and Jordan, T. R. (2004) Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 30: 873–88.

van Wassenhove, V. (2004) Cortical dynamics of auditory-visual speech: a forward model of multisensory integration. Doctoral dissertation, University of Maryland.

Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., and Munhall, K. G. (1998) Eye movement of perceivers during audiovisual speech perception. *Perception and Psychophysics*, 60: 926–40.

Vitkovich, M., and Barber, P. (1994) Effects of video frame rate on subjects' ability to shadow one of two competing verbal passages. *Journal of Speech and Hearing Research*, 37: 1204–10.

Vroomen, J., Driver, J., and de Gelder, B. (2001) Is cross-modal integration of emotional expressions independent of attentional resources? *Cognitive and Affective Neurosciences*, 1: 382–7.

Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., and Jones, C. J. (1977) Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research*, 20: 130–45.

Walden, B. E., Prosek, R. A., and Worthington, D. W. (1974) Predicting audiovisual consonant recognition performance of hearing-impaired adults. *Journal of Speech and Hearing Research*, 17: 270–78.

Williams, J. H., Whiten, A., Suddendorf, T., and Perrett, D. I. (2001) Imitation, mirror neurons and autism. *Neuroscience and Biobehavior Review*, 25: 287–95.

Wozniak, V. D., and Jackson. P. L. (1979) Visual vowel and diphthong perception from two horizontal viewing angles. *Journal of Speech and Hearing Research*, 22: 354–65.

Yehia, H. C., Rubin, P. E., and Vatikiotis-Bateson, E. (1998) Quantitative association of vocal-tract and facial behavior. *Speech Communication*, 26: 23–44.

Zwitserlood, P. (1989) The locus of the effects of sentential-semantic context in spoken-word processing. *Cognition*, 32: 25–64.