9.1 Introduction

The seemingly effortless and incidental way in which infants acquire and perfect the ability to use spoken language/s is quite remarkable considering the complexity of human linguistic systems. Parsing the spoken signal successfully is in fact anything but trivial, as attested by the challenge posed by learning a new language later in life, let alone attaining native-like proficiency. It is now widely accepted that (adult) listeners tend to make use of as many cues as there are available to them in order to decode the spoken signal effectively (e.g. Cutler 1997; Soto-Faraco et al. 2001). Different sources of information not only span various linguistic levels (acoustics, phonology, lexical, morphology, syntax, semantics, pragmatics, etc.), but also encompass different sensory modalities such as vision and audition (Campbell et al. 1998) and even touch (see Gick and Derrick 2009). Thus, like most everyday life perceptual experiences (Gibson 1966; Lewkowicz 2000a; Stein and Meredith 1993), spoken communication involves multisensory inputs. In the particular case of speech, the receiver in a typical face-to-face conversation has access to the sounds as well as to the corresponding visible articulatory gestures made by the speaker. The integration of heard and seen speech information has received a good deal of attention in the literature, and its consequences have been repeatedly documented at the behavioral (e.g. Ma et al. 2009; McGurk and MacDonald 1976; Ross et al. 2007; Sumby and Polack 1954) as well as the physiological (e.g. Calvert et al. 2000) level. The focus of the present chapter is on the developmental course of multisensory processing mechanisms that facilitate language acquisition, and on the contribution of multisensory information to the development of speech perception. We argue that the plastic mechanisms leading to the acquisition of a language, whether in infancy or later on, are sensitive to the correlated and often complementary nature of multiple crossmodal sources of linguistic information. These crossmodal correspondences are used to decode and represent the speech signal from the first months of life.

9.2 Early multisensory capacities in the infant

From a developmental perspective, the sensitivity to multisensory redundancy and coherence is essential for the emergence of adaptive perceptual, cognitive, and social functioning (Bahrick et al. 2004; Gibson 1969; Lewkowicz and Kraebel 2004; Spelke 1976; Thelen and Smith 1994). This includes the ability to extract meaning from speech, arguably the most important communicative signal in an infant’s life. The multisensory coherence of audiovisual speech is determined by the overlapping nature of the visual and auditory streams of information emanating from the face and vocal tract of the speaker, reflected in a variety of amodal stimulus attributes (such as
intensity, duration, tempo, and rhythm) and temporally and spatially contiguous modality-
specific attributes (such as the color of the face and the pitch and timbre of the voice). A priori, it
is reasonable to assume that an immature and inexperienced infant should not be able to extract
the intersensory relations offered by all these cues. This assumption is consistent with the classic
developmental integration view, according to which, initially in life, infants are not capable of
perceiving crossmodal correspondences and only acquire this ability gradually as a result of expe-
rience (Birch and Lefford 1963, 1967; Piaget 1952). However, this idea stands in contrast to the
other classic developmental framework, namely the developmental differentiation view (Gibson
1969, 1984), according to which a number of basic intersensory perceptual abilities are already
present at birth (although which specific ones are present has never been specified). Furthermore,
the differentiation view assumes that as infants discover increasingly more complex unsensory
features of their world through the process of perceptual learning and differentiation they can
also discover increasingly more complex types of intersensory relations.

A great deal of empirical work on the development of intersensory perception has accumulated
since these two opposite theoretical frameworks were first proposed and it has now become abun-
dantly clear that both developmental integration and differentiation processes contribute to the
emergence of intersensory processing skills in infancy (Lewkowicz 2002). On the one hand, it has
been found that infants exhibit some relatively primitive intersensory perceptual abilities, such as
the detection of synchrony and intensity relations early in life (e.g. Lewkowicz 2010; Lewkowicz
et al. 2010; Lewkowicz and Turkewitz 1980). On the other, it has been found that infants gradu-
ally acquire the ability to perceive more complex types of intersensory relations as they acquire
perceptual experience (Lewkowicz 2000a, 2002; Lewkowicz and Ghazanfar 2009; Lewkowicz and
Lickliter 1994; Lickliter and Bahrick 2000; Walker-Andrews 1986, 1997). Finally, it is worth
pointing out that, just like in other perceptual domains (e.g. Werker and Tees 1984), the way
in which intersensory development progresses is not always incremental (a trend that will be
discussed in the following sections). For example, recent studies have shown that young infants
can match other-species’ faces and vocalizations (Lewkowicz et al. 2010; Lewkowicz and Ghazanfar
2006) or non-native visible and audible speech events (Pons et al. 2009; see Section 9.5) but when
they are older infants no longer make intersensory matches of non-native auditory and visual
inputs (i.e. those from other species or from a non-native language).

The balance between the early intersensory capabilities on the one side and the limitations due
to the immaturity of the nervous system and relative perceptual inexperience on the other, raises
several questions regarding infants’ initial perception of audiovisual speech.

◆ At what point in early development does the ability to perceive the multisensory coherence of
audiovisual speech emerge?
◆ When this ability emerges, what specific attributes permit its perception as a coherent event?
◆ Do certain perceptual attributes associated with audiovisual speech have developmental
priority over others?

The answers to these questions depend, first and foremost, on an analysis of the audiovisual
speech signal. In essence, audiovisual speech is represented by a hierarchy of audiovisual relations
(Munhall and Vatikiotis-Bateson 2004). These consist of the temporally synchronous and spa-
tially contiguous onsets and offsets of facial gestures and vocalizations, the correlation between
the dynamics of vocal-tract motion and the dynamics of accompanying vocalizations (that are
usually specified by duration, tempo, and rhythmical pattern information), and finally by various
amodal categorical attributes such as the talker’s gender, affect, and identity. Whether infants can
perceive some or all of these different intersensory coherence cues is currently an open question,
especially during the early months of life.
EARLY PERCEPTUAL SENSITIVITY TO HEARD AND SEEN SPEECH

So far, a number of studies are consistent with the idea that infants can perceive coherence between isolated auditory and visual phonemes from as young as 2 months of age (Kuhl and Meltzoff 1982, 1984; Patterson and Werker 1999, 2003; Pons et al. 2009; though see Walton and Bower 1993, for potential evidence of matching at even younger ages). Moreover, some researchers have attempted to extrapolate audiovisual matching abilities to longer speech events (e.g. Dodd 1979, 1987; Dodd and Burnham 1988). However, it is still unclear what particular sensory cues infants respond to in these studies. More recently, it has been reported that by around 4 months of age infants exhibit the McGurk effect (Burnham and Dodd 2004; Desjardins and Werker 2004; Rosenblum et al. 1997; see Section 9.3), suggesting that infants, like adults, can integrate incongruous audible and visible speech cues. Finally, it has been reported that infant discrimination of their mother’s face from that of a stranger is facilitated by the concurrent presentation of the voice, indicating the ability to use crossmodal redundancy (e.g. in 4–5-months olds; Burnham 1993, and even newborns; Sai 2005).

The findings regarding infants’ abilities to match visible and audible phonemes are particularly interesting because at first glance they seem to challenge the notion that early intersensory perception is based on low-level cues such as intensity and temporal synchrony. Indeed, with specific regards to temporal synchrony, some of the studies to date have demonstrated phonetic matching across auditory and visual stimuli presented sequentially rather than simultaneously (therefore, in the absence of crossmodal temporal synchrony cues; e.g. Pons et al. 2009). As a result, the most reasonable conclusion to draw from these studies is that infants must have extracted some higher-level, abstract, intersensory relational cues in order to perform the crossmodal matches (i.e. the correlation between the dynamics of vocal-tract motion and the dynamics of accompanying vocalizations; see MacKain et al. 1983, for a similar argument). As indicated earlier, audiovisual speech is a dynamic and multidimensional event that can be specified by synchronous onsets of acoustic and gestural energy as well as by correlated dynamic patterns across audible and visible articulations (Munhall and Vatikiotis-Bateson 2004; Yehia et al. 1998). Thus, the studies showing that infants can match faces and voices across a temporal delay show that they can take advantage of one perceptual cue (e.g. correlated dynamic patterns) in the absence of the other (e.g. synchronous onsets). This is further complemented by findings that infants can also perceive audiovisual synchrony cues in the absence of higher-level intersensory correlations (Lewkowicz 2010; Lewkowicz, et al. 2010). Therefore, when considered together, current findings suggest that infants are sensitive to both correlated dynamic patterns of acoustic and gestural energy as well as to synchronous onsets, and that they can respond to each of these intersensory cues independently.

9.3 Early perceptual sensitivity to heard and seen speech

Previous research has revealed that, from the very first moments in life, human infants exhibit sensitivity to the perceptual attributes of speech, both acoustic and visual. For example, human neonates suck preferentially when listening to speech versus complex non-speech stimuli (Voulumanos and Werker 2007). As mentioned in the previous section, this sensitivity can be fairly broad in the newborn (e.g. extending both to human speech and to non-human vocalizations), but within the first few months of life it becomes quite specific to human speech (Vouloumanos et al. 2010). Moreover, human neonates can acoustically discriminate languages from different rhythmical classes (Byers-Heinlein et al. 2010; Mehler et al. 1988; Nazzi et al. 1998) and show categorical perception of phonetic contrasts (Dehaene-Lambertz and Dehaene 1994). At birth, infants can discriminate both native and unfamiliar phonetic contrasts, suggesting that the auditory/linguistic perceptual system is broadly sensitive to the properties of any language.
Then, over the first several months after birth, phonetic discrimination becomes reorganized resulting in an increase in sensitivity to the speech contrasts in the infant’s native language (Kuhl et al. 2003, 2006; Narayan et al. 2010) and a concomitant decrease in their sensitivity to those phonetic contrasts that are not used to distinguish meaning in the infant’s native language (Werker and Tees 1984; see Safran et al. 2006 for a review).

The perceptual sensitivity to the attributes of human speech early in life is paralleled by specialized neural systems for the processing of both the phonetic (e.g. Dehaene-Lambertz and Baillet 1998; Dehaene-Lambertz and Dehaene 1994) and rhythmical (e.g. Dehaene et al. 2006; Peña et al. 2003) aspects of spoken language. Indeed, some studies indicate that newborn infants are sensitive to distinct patterns of syllable repetition (i.e. they can discriminate novel speech items following an ABB syllabic structure—e.g. wefofo, gubibi, etc—from those following an ABC structure). This discriminative ability seems to implicate left hemisphere brain areas analogous to those involved in language processing in the adult brain (Gervain et al. 2008) and dissociable from areas implicated in tone sequence discrimination (Gervain et al. 2009). This differential pattern of neural activation suggests that infants’ sensitivity to speech-like stimuli may involve early specialization.

Although visual language discrimination has not been studied in the newborn (in the same way that auditory language has), there is evidence that infants as young as 4 months are already able to discriminate languages just by watching silent talking faces (Weikum et al. 2007). In the study by Weikum and colleagues, infants were presented with videos of silent visual speech in English and French in a habituation paradigm (see Fig. 9.1). Monolingual infants of 4 and 6 months revealed sensitivity to a language switch (in the form of increased looking times), but appeared to lose this sensitivity at 8 months of age. On the contrary, 8-month olds raised in a bilingual environment maintained the sensitivity to switches between their two languages. Thus, the sensitivity to visual speech cues begins to decline by 8 months of age, unless the infant is being raised in a bilingual environment. It would be interesting to address whether this early visual sensitivity to distinct languages exploits rhythmical differences (as seems to be the case for auditory discrimination), dynamic visual phonetic (i.e. visegmetic) information (which differs between English and French), or both.

Young infants are not only sensitive to the auditory and visual concomitants of speech in isolation, they are also sensitive, from a very early age, to the percepts that arise from the joint influence of both auditory and visual information. For example, as mentioned in the previous section, the presence of the maternal voice has been found to facilitate recognition of the maternal face in newborn infants (Sai 2005) as well as in 4–5-month-old infants (Burnham 1993; Dodd and Burnham 1988). These findings suggest that sensitivity to speech, evident early in life, can support and facilitate learning about specific faces. Perhaps even more remarkable is the extent to which young infants are sensitive to matching information between heard and seen speech (Baier et al. 2007; Kuhl and Meltzoff 1982, 1984; MacKain et al. 1983; Patterson and Werker 1999, 2002; Yeung and Werker, in preparation). For example, in Kuhl and Meltzoff’s study, 2- and 4-month-old infants looked preferentially at the visual match when shown side-by-side video displays of a woman producing either an ‘ee’ or an ‘ah’ vowel together with the soundtrack of one of the two

---

1 The term ‘visegmetic’ is used here to refer to the dynamically changing arrangement of features on the face (mouth, chin, cheeks, head etc.). In essence, it refers to anything visible on the face and head that accompanies the production of segmental phonetic information. Visegmetic segments are therefore meant to capture all facets of motion that the face and head make as language sounds are produced, and include the time-varying qualities that are roughly equivalent to auditory phonetic segments.
vowels. Some authors have reported that, even at birth, human infants seem to show an ability to match human facial gestures with their corresponding heard vocalizations (Aldridge et al. 1999). Yet, although it appears that there is substantial organization from an early age that supports intersensory matching, this ability is amenable to experience-related changes. For example, following a cochlear implant in infants born deaf, the propensity for intersensory matching of

---

**Fig. 9.1** (a) Schematic representation of the habituation paradigm used in Weikum et al.’s (2007) visual language discrimination experiment. Infants (4, 6 and 8 months old) were exposed to varied speakers (only one shown in the figure) across different trials, but all could speak either English or French. Some example snap-shots of the set up are presented in the top illustrations. Infants sat in their mother’s lap. She was wearing darkened glasses to prevent her from seeing the video-clips the infant was being presented with. (b) Results (in average looking times) are presented separately from the different age groups (columns) and linguistic backgrounds (rows) tested in the experiment. In each graph, the looking times for the final phase the habituation period and the test phase are presented for the language switch and no-switch (control) trials. Error-bars represent the SEM.
speech begins returning within weeks and months (Barker and Tomblin 2004; see Section 9.7). Furthermore, during the first year of life, intersensory matching becomes increasingly attuned to the native input of the infant (Pons et al. 2009).

The intersensory matching findings discussed so far highlight the remarkable ability of infants to link phonetic characteristics across the auditory and visual modality. However, natural speech processing requires the real-time integration of concurrent heard and seen aspects of speech. One of the most interesting illustrations of real-time audiovisual integration in adult speech processing is that visible speech can change the auditory speech percept in the case of crossmodal conflict (McGurk and MacDonald 1976). Several studies have indicated that this illusion, known as the McGurk effect, is present as early as 4 months of age, thus implying that infants are able to integrate seen and heard speech (e.g. Burnham and Dodd 2004; Desjardins and Werker 2004; Rosenblum et al. 1997). For example, in Burnham and Dodd’s study, infants were habituated to either a congruent auditory /ba/ + visual [ba] or an incongruent auditory /ba/ + visual [ga]; the latter leads to a ‘da’ or ‘tha’ percept in adults. After habituation, the infants’ looking times to each of three different auditory-only syllables, /ba/, /da/, and /tha/, all of which were paired with a still face, were measured. Habituation to the audiovisually congruent syllable /ba/ + [ba] made infants look longer to /da/ or /tha/ than to /ba/, indicating that infants perceived /ba/ as different from the other two syllables. Furthermore, and consistent with integration, after habituation to the incongruent audiovisual stimulus (auditory /ba/ + visual [ga]) infants looked longer to /ba/, implying that they perceived this syllable as a novel stimulus. Other results confirm the finding of integration from an early age using different stimuli, albeit not in every combination in which it should have been predicted from adult performance (Desjardins and Werker 2004; Rosenblum et al. 1997; see Kushnirenko et al. 2008, discussed in Section 9.6 for electrophysiological evidence).

Beyond the illusions arising from crossmodal conflict, pairing congruent visual information and speech sounds can facilitate phonetic discrimination in infancy (Teinonen et al. 2008; see Navarra and Soto-Faraco 2007 for evidence in adults). In Teinonen et al.’s study, infants aged 6 months were familiarized to a set of acoustic stimuli selected from a restricted range in the middle of the /ba/–/da/ continuum (and thus, somehow acoustically ambiguous), paired with a visual display of a model articulating either a canonical [ba] or a canonical [da]. In the consistent condition, auditory tokens from the /ba/ side of the boundary were paired with visual [ba] and all audio /da/ tokens were paired with visual [da]. In the inconsistent condition, audio and visual stimuli were randomly paired. Only infants in the consistent pairing condition were able to discriminate /ba/ versus /da/ stimuli from either side of the boundary in a subsequent auditory

---

2 Referred to as the McGurk effect or McGurk illusion, when adults are presented with certain mismatching combinations of acoustic and visual speech tokens (such as an auditory /ba/ in synchrony with a seen [ga]), the resulting percept is fusion between what is heard and seen (i.e. in this example it would result in a perceived ‘da’ or ‘tha’, an entirely illusory outcome). Intersensory conflict in speech can also result in combinations rather than fusions (for example, when presented with an auditory /ga/ and a visual [ba], adults often report perceiving ‘bga’) and in visual capture (for example, when presented with a heard /ba/ and a seen [va], participants often report hearing ‘va’). Even when aware of the dubbing procedure and the audiovisual mismatch, adults cannot inhibit the illusion, though some limitations due to attention (Alsius et al. 2005, 2007; Navarra et al. 2010, for a review) and experience (Sekiyama and Burnham 2008; Sekiyama and Tohkura 1991, 1993, see also Section 9.7) have been reported (some have even argued that it is a learned phenomenon; e.g. Massaro 1984). The existence of this illusory percept arising from conflict between seen and heard speech in infancy provides much stronger evidence of integration than what is shown by bimodal matching.
discrimination test. Teinonen et al.’s results provide support for the notion that visual and auditory speech involve similar representations, and that their co-occurrence in the world might support learning of native category boundaries. However, this phenomenon might extend to other types of correlations.

In a recent study, Yeung and Werker (2009) found that simply pairing auditory speech stimuli with two different (non-speech) visual objects also facilitates discrimination. Yeung and Werker tested 9-month-old infants of English-speaking families using a retroflex /Da/ versus a dental /da/, a Hindi contrast, which does not distinguish meaning in English. Prior studies have already shown that English infants are sensitive to this contrast at 6–8 months of age, but not at 10–12 months (Werker and Lalonde 1998; Werker and Tees 1984). Yeung and Werker demonstrated, however, that exposure to consistent pairings of /da/ with one visual object and /Da/ with another during the familiarization phase facilitated later discrimination of this non-native contrast, whereas following an inconsistent pairing, the 9-month-old English infants could not discriminate. Yeung and Werker’s results raise the possibility that it is not the matching between specific audiovisual linguistic information that enhanced performance in Teinonen et al.’s study, but rather that it is just the contingent pairing between phonemic categories and visual objects.

### 9.4 Developmental changes in audiovisual speech processing from childhood to adulthood

Despite the remarkable abilities of infants to extract and use audiovisual correlations (in speech and other domains) from a very early age (e.g. Burnham 1993; Kuhl and Meltzoff 1982, 1984; Lewkowicz 2010; Lewkowicz et al. 2010; Lewkowicz and Turkewitz 1980; MacKain et al. 1983; Patterson and Werker 1999, 2003; Sai 2005; Walton and Bower 1993), it also becomes clear from the literature reviewed above that there are very important experience-related changes in the development of audiovisual speech processing (e.g. Barker and Tomblin 2004; Lewkowicz 2000a, 2002; Lewkowicz and Ghazanfar 2006, 2009; Lewkowicz and Lickliter 1994; Lewkowicz, et al. 2010; Lickliter and Bahrick 2000; Pons et al. 2009; Teinonen et al. 2008; Walker-Andrews 1997). These changes, however, do not stop at infancy but continue throughout childhood and beyond. A good illustration of this point is provided in the paper where the now famous McGurk effect was first reported (McGurk and MacDonald 1976). McGurk and MacDonald’s study was in fact a cross-sectional investigation including children of different ages (3–5 and 7–8 year-olds) as well as adults (18–40 years old), and the results pointed out a clear discontinuity between middle childhood and adulthood. In McGurk and MacDonald’s study, children of both age groups were less amenable to the audiovisual illusion than adults, a consistent pattern that has been replicated in subsequent studies (e.g. Desjardins et al. 1997; Hockley and Polka 1994; Massaro 1984; Massaro et al. 1986; Sekiyama and Burnham 2008). This developmental trend is usually accounted for by the reduced amount of experience with (visual) speech in children as well as, according to some authors, by the reduced degree of sensorimotor experience resulting from producing speech (e.g. Desjardins et al. 1997). Massaro et al. pointed out that rather than reflecting a different strategy of audiovisual integration in children, the reduced visual influence at this age range was better accounted for by a reduction in the amount of visual information extracted from the stimulus. Consistent with this interpretation, children seem to be poorer lip-readers than adults (Massaro et al. 1986). The fact that young infants exhibit the McGurk illusion at all (see Section 9.3) and that they are able to extract sufficient visual information to discriminate between languages (Weikum et al. 2007) seems to indicate, however, that visual experience may not be the only factor at play in the development of the integrative mechanisms, or at the least that a minimal
amount of experience is sufficient to set up the rudiments of audiovisual integration (albeit not up to adult-like levels).

This general trend of increasing visual influence in speech perception toward the end of childhood seems to be modulated by environmental factors, such as the language background and/or cultural aspects. This comes across clearly in the light of the remarkable differences in audiovisual speech perception between different linguistic groups in adulthood. For example, Japanese speakers are less amenable to the McGurk illusion and to audiovisual speech integration in general than westerners (e.g., Sekiyama and Burnham 2008; Sekiyama and Tohkura 1991, 1993; see also Sekiyama 1997, for Chinese). The accounts of this cross-linguistic difference in susceptibility to audiovisual effects have varied. One initial explanation pointed at cultural differences in terms of how much people tend to look at each other’s faces while speaking. In particular, direct gaze is less frequent in Japanese society (e.g. Sekiyama and Tohkura 1993), leading to a smaller amount of visual experience with speech and, as a consequence, reducing its contribution in audiovisual speech perception. Nevertheless, not all cross-linguistic differences in audiovisual speech processing may be so easily explained along such cultural lines (e.g. Aloufy et al. 1996 for reduced visual influence in Hebrew versus English speakers). Other, more recent, explanations for differences in audiovisual speech integration across languages focus on how the phonological space of different languages may determine the relative reliance on the auditory versus visual inputs by their speakers. Because the phonemic repertoire of Japanese is less crowded than that of English, Japanese speakers would require less visual aid in everyday spoken communication thereby learning to rely less strongly in the visual aspects of speech (e.g. Sekiyama and Burnham 2008).

Another possibly important source of change in audiovisual speech integration during adulthood relates to variations in sensory acuity throughout life. Along these lines, it has been claimed that multisensory integration in general becomes more important in old age, because the general decline in unisensory processing would render greater benefits of the integration between different sensory sources as per the rule of inverse effectiveness (Laurienti et al. 2006; also see Chapter 11). However, several previous studies testing older adults have failed to meet the prediction of increased strength of audiovisual speech integration as compared to younger adults (Behne et al. 2007; Cienkowski and Carney 2002; Middleweerd and Plomp 1987; Sommers et al. 2005; Walden et al. 1993). In fact, some of these studies reported superior audiovisual integration in adults of younger age when differences in hearing thresholds are controlled for (e.g. Sommers et al. 2005). In a similar vein, Ross et al. (2007) reported a negative correlation between age and audiovisual performance levels across a sample ranging from 18 to 59 years old. The finding of reduced audiovisual integration in older versus younger adults could be caused by a decline in unisensory (visual speech reading) in the elderly (e.g. Arlinger 1991; Gordon and Allen 2009). At present, the conflict between studies showing increased audiovisual influence with ageing, and those reporting no change or even decline is unresolved.

9.5 Development of audiovisual speech processing in multilingual environments

An often overlooked fact is that, for a large part of the world’s population, language acquisition takes place in a multilingual environment. However, the role of audiovisual processes in the context of multiple input languages has received comparatively little attention from researchers (see Burnham 1998 for a review). One of the initial challenges posed for infants born in a multilingual environment is undoubtedly the necessity to discriminate between the languages spoken around them (Werker and Byers-Heinlein 2008). From a purely auditory perspective, current evidence
highlights the key role played by prosody in language discrimination. For example, the ability to
differentiate between languages belonging to different rhythmic classes (e.g. stress-timed lan-
guages, such as English, versus syllable-timed languages, such as Spanish; see Mehler et al. 1988)
is present at birth whereas the ability to discriminate between languages belonging to the same
rhythmic class emerges by 4–5 months of age (Bosch and Sebastian-Galles 2001; Nazzi and
Ramus 2003). Moreover, infants as young as 4 months of age can discriminate languages on the
basis of visual information (Weikum et al. 2007, described earlier). This last finding raises the
interesting possibility, mentioned earlier, that, just as in infant auditory language discrimination,
rhythmic attributes play an important role in infant visual language discrimination. This is made
all the more likely by the fact that rhythm is an amodal stimulus property to which infants are
sensitive visually as well as acoustically from a very early age (Lewkowicz 2003; Lewkowicz and
Marcovitch 2006) and by the fact that rhythmic properties seem to play an important role in
visual language discrimination in adults (Soto-Faraco et al. 2007). For instance, in Soto-Faraco
et al. (2007), one of the few variables that correlated with (Spanish–Catalan) visual language dis-
crimination ability in adults was vowel-to-consonant ratio, a property that is related to rhythm.
One should nevertheless also consider the role of visegmetic (visual phonetic) information as a
cue to visual language discrimination (Weikum et al. in preparation).

What about matching languages across sensory modalities? One of the few studies addressing
audiovisual matching cross-linguistically was conducted by Dodd and Burnham (1988). The
authors measured orienting behaviour in infants raised in monolingual English families when
presented with two faces side by side, one silently articulating a language passage in Greek and the
other another passage in English, accompanied by either the appropriate Greek or the appropriate
English passage played acoustically. Infants of 2½ months did not show significant audiovisual
language matching in either condition whilst 4½-month-olds significantly preferred to look at
the English-speaking face while hearing the English soundtrack. Unfortunately, given that there
was an overall preference for the native-speaking faces, it is unclear whether these results reflect
true intersensory matching or a general preference for the native language.

Phoneme perception and discrimination behaviour offers a further opportunity to study
mono- and multilingual aspects of audiovisual speech development. A pioneering study by
Werker and Tees (1984) using acoustic stimuli showed that the initial broad sensitivity to native
and non-native phonemic contrasts by 6 months of age narrowed down to only those contrasts
that are relevant to the native language by 12 months of age. This type of phenomenon, referred
to as perceptual narrowing, has recently been shown to occur in visual development (i.e. in the
discrimination of other-species faces; Pascalis et al. 2005), in matching of other-species faces and
vocalizations (Lewkowicz and Ghazanfar 2006, 2009), and, interestingly, in matching auditory
phonemes and visual speech gestures from a non-native language (Pons et al. 2009). In particular,
the study by Pons et al. revealed that narrowing of speech perception to the relevant contrasts in
the native language is not limited to one modality (acoustic or visual), but that it is indeed a
pansensory\(^3\) and domain-general phenomenon. Using an intersensory matching procedure, Pons
and colleagues familiarized infants from Spanish and English families with auditory /ba/ or /va/, an
English phonemic contrast which is allophonic in Spanish (see Fig. 9.2). The subsequent test
consisted of side-by-side silent video clips of the speaker producing each of these syllables.
Whereas 6-month-old Spanish-learning infants looked longer at the silent videos matching the
pre-exposed syllable, 11-month-old Spanish-learning infants did not, indicating that their ability
to match the auditory and visual representation of non-native sounds declines during the first

\(^3\) ‘Pansensory’ is used here in the sense that it applies to all sensory modalities.
year of life. In contrast, English-learning infants, whose phonological environment includes both /ba/ and /va/, exhibited intersensory matching at both ages. Interestingly, once the decline occurs it seems to persist into adulthood, given that Spanish-speaking adults (tested on a modified intersensory matching task) exhibited no evidence of intersensory matching for /va/ and /ba/, whereas English-speaking adults exhibited matching accuracy above 90%.

Fig. 9.2 (a) Schematic representation of the preferential looking paradigm used in Pons et al.’s (2009) audiovisual matching experiment. Infants preferential looking to each silent vocalization was measured before (baseline) and after (test) a purely acoustic familiarization period where one of the two syllables was exposed to the infant. During the familiarization phase, a rotating circle was shown at the centre of the screen. (b) Results (in relative preferential looking times) are presented separately from the different infant age groups and linguistic backgrounds tested in the experiment. Empty circles represent individual scores, and filled dots represent group average. (c) Spanish adult results (proportion correct) for the phonemic discrimination procedure; other conventions are as in (b).
Evidence from the cross-linguistic studies discussed earlier (Section 9.4) suggests that adult audiovisual speech processing can be highly dependent on the language-specific attributes experienced throughout development (e.g. Sekiyama 1997; Sekiyama and Tohkura 1991, 1993). To track the developmental course of these differences, Sekiyama and Burnham (2008) tested 6-, 8- and 11-year-old English and Japanese children as well as adults in a speeded classification task for unimodal and crossmodal (congruent and incongruent) /ba/, /da/, and /ga/ syllables. Despite the equivalent degree of visual interference from audiovisually incongruent syllables of the two language groups at 6 years of age, from 8 years onward the English participants were more prone to audiovisual interference than the Japanese participants. The authors of this study speculated that the reasons for this difference may be due to the complexity of English phonology showing a larger benefit of redundant and complementary information from the visual modality relative to the comparatively less crowded Japanese phonological space. Sekiyama and Burnham proposed that this would be exacerbated with the commencement of school by 8 years of age as children enter in contact with larger talker variability. In a related study, Chen and Hazan (2007) tested the degree of visual influence on auditory speech in children between 8 to 9 years of age and adult participants in monolingual (English and Mandarin native speakers) and one early bilingual group (children only, Mandarin natives who were learning English). The stimuli were syllables presented in auditory, visual, crossmodal congruent, and crossmodal incongruent (McGurk-type) conditions. Although the data revealed a developmental increase in visual influence (in accord with much of the previous literature), in contrast to Sekiyama and Burnham (2008) no differences in the size of visual influence as a function of language background or bilingual experience were found.

Finally, recent work by Kuhl et al. (2006) provides some further insights into the limits and plasticity of crossmodal speech stimulation for perceiving non-native phonological contrasts. Kuhl et al. (2006) reported two studies in which 9-month old infants raised in an American English language environment were exposed to Mandarin Chinese for twelve 25-minute sessions over a month, either in a social (face-to-face) situation or through a pre-recorded presentation played back on a TV set. The post-exposure discrimination tests showed that only those infants in the face-to-face sessions were able to perceive Mandarin phonetic contrasts at a similar level to native Mandarin Chinese infants. These infants clearly outperformed a control group who underwent similar face-to-face exposure sessions but in (their native) English language. This finding was interpreted as indicating that social context is necessary for maintaining or re-gaining sensitivity to non-native phonetic contrasts by the end of the first year of life. However, the actual role of increased attention associated with personal interactions, as compared to watching a pre-recorded video-clip, should also be considered. That is, by 9 months of age children still retain some residual flexibility for non-native contrast discrimination, which can be maintained simply by increasing passive exposure time (from 2 min at 6–8 months; Maye et al. 2002, to 4 min of exposure at 10 months; Yoshida et al. 2010). Thus although social interaction is effective in facilitating the relearning of non-native distinctions after 9 months of age, it may be so not because it is social per se, but because it increases infant attention by providing a richly referential context in which objects are likely being labelled when the infant is paying attention.

9.6 Neural underpinnings of the development of audiovisual speech processing

Only a handful of studies have dealt with the neural underpinnings of audiovisual speech perception in infants. One behavioral study that is worth mentioning is the one conducted by MacKain et al. (1983), who first suggested the lateralization of audiovisual matching of speech to the left
cerebral hemisphere in infancy. In this study, 5–6-month-old infants were presented with an auditory sequence of disyllabic pseudo-words (e.g. /vava/, /zuzu/...) together with two side-by-side video-clips of a speaking face. Both speaking faces uttered two syllable pseudo-words, temporally synchronized with the syllables presented acoustically, but only one of the visual streams was congruent in segmental content with the auditory stream. Infants preferred to look selectively at the video-clip which was linguistically congruent with the soundtrack, but only if it was shown on the right side of the infant’s field of view. This finding resonates well with the existence of an early capacity for audiovisual matching and integration within the first few months of life revealed by other studies (Baier et al. 2007; Burnham and Dodd 2004; Desjardins and Werker 2004; Kuhl and Meltzoff 1982; 1984; Rosenblum et al. 1997). In McKain et al. ’s study, owing to the lateralization of gaze behaviour to the contralateral hemispace in tasks engaging one hemisphere more than the other, the authors inferred left-hemisphere specialization for audiovisual speech matching. This would suggest an early functional organization of the neural substrates subserving audiovisual speech processing.

There are only a few studies that have directly related physiological measurements in infants to crossmodal speech perception. Kushnerenko and colleagues (2008) recorded electrical activity from the brains of 5-month-old infants while they listened to and watched the syllables /ba/ and /ga/. For each stimulus, the modalities could be crossmodally matched or mismatched in content. In one of the mismatched conditions where auditory /ba/ was dubbed onto visual [ga], the expected perceptual fusion (‘da’, as per adult and infant results, see Burnham and Dodd 2004) does not violate phonotactics in the infant’s native language (i.e. English). In this case, the electrophysiological trace was no different from the crossmodally matching stimuli (/ba/ + [ba], or /ga/ + [ga]). Crucially, in the other mismatched condition auditory /ga/ was dubbed onto visual [ba], a stimulus perceived as ‘bga’ by adults and therefore in violation of English phonotactics. In this case, the trace evoked by the crossmodal mismatch was significantly different from the matching crossmodal stimuli (as well as from the other crossmodal mismatch leading to phonotactically legal ‘da’). The timing and scalp distribution of the differences in evoked potentials suggested that the effects of audiovisual integration could be being reflected in the auditory cortex. The expression of audiovisual integration at early sensory stages coincides with what is found in adults (e.g. Sams et al. 1991), and therefore Kushnerenko et al. ’s results suggest that the organization of adult-like audiovisual integration networks starts early on in life.

The results of Bristow and colleagues (2008) with 10-week old infants further converge on the conclusion that the brain networks supporting audiovisual speech processing are already present, albeit at an immature stage, soon after birth. In their experiment, infants’ mismatch responses in evoked potentials were measured to acoustically presented vowels (e.g. /a/) following a brief habituation period to the same or a different vowel (e.g. /i/). Mismatch responses were very similar regardless of the sensory modality in which the vowels had been presented during the habituation period (auditory or visual). This led Bristow et al. to conclude that infants at this early age have crossmodal or amodal phonemic representations, in line with the results of behavioural studies revealing an ability to match speech intersensorially (e.g. Kuhl and Meltzoff 1982, 1984; Patterson and Werker 2003). Interestingly, the scalp distribution and the modelling of the brain sources of the mismatch responses revealed a left lateralized cortical network responsive to phonetic matching that is dissociable from areas sensitive to other types of crossmodal matching (such as face–voice–gender). Remarkably, at only 10 weeks of age, this network already has substantial overlap with audiovisual integration areas typically seen in adults, including frontal (left inferior frontal cortex) as well as temporal regions (left superior and left inferior temporal gyrus).

Recent work from Dick and colleagues (Dick et al. 2010) sheds some light on the brain correlates of audiovisual speech beyond infancy (~9 years old). This is an interesting age range because
it straddles the transition period between the reduced audiovisual integration levels in infancy and adult-like audiovisual processing, discussed in Section 9.4 (Desjardins et al. 1997; Hockley and Polka 1994; Massaro 1984; Massaro et al. 1986; McGurk and MacDonald 1976; Sekiyama and Burnham 2008). Dick et al. demonstrated that by nine years of age children rely on the same network of brain regions used by adults for audiovisual speech processing. However, the study of the information flow with structural equation modelling demonstrated differences between adults and children (the posterior inferior frontal gyrus/ventral premotor cortex modulations on supramarginal gyrus differed across age groups) and these differences were attributed to less efficient processing in children. The authors interpreted these findings as the result of maturational processes modulated by language experience.

A clear message emerges in the light of the physiological findings discussed so far; the basic brain mechanisms that will support audiovisual integration in adulthood begin to emerge quite early in life, albeit in an immature form. Thus these mechanisms and their supporting neural substrates seem to undergo a slow developmental process from early infancy up until the early stages of adulthood. This fits well with previous claims that the integration mechanisms in infants and adults are not fundamentally different, and that it is the changes in the amount of information being extracted from each input that determines the output of integration throughout development (Massaro et al. 1986). In addition, the seemingly long period of time that it takes before adult-like levels of audiovisual processing are achieved suggests that there are lots of opportunities for experience to continually contribute to the development of these circuits.

9.7 Development of audiovisual speech after sensory deprivation

As many as 90–95% of deaf children are born to hearing parents and thus they grow up in normal-hearing linguistic environments where they experience the visual correlates of spoken language (see Campbell and MacSweeney 2004). What is more, in many cases visual speech becomes the core of speech comprehension. A critical question is thus whether speech-reading, or lip-reading, can be improved with training and, related, whether deaf people are better speech-readers than hearing individuals by way of their increased practice with visual speech. On the one hand, several studies have shown that a percentage of individuals born deaf have better lip-reading abilities than normal-hearing individuals (see Auer and Bernstein 2007; Bernstein et al. 2000; Mohammed et al. 2005), although others have failed to find that deaf people are superior lip readers (see Mogford 1987). When looking at the effectiveness of specific training in lip-reading, success is very limited, at least with short training periods (e.g. 5 weeks), in both hearing individuals as well as in inaccurate deaf lip-readers (Bernstein et al. 2001), suggesting that longer-term experience may be necessary for the achievement of lip-reading skills. On the other hand, the differences between congenital and developmentally deaf individuals reveal that one key factor for the achievement of effective speech-reading abilities is early experience with the correspondence between auditory and visual speech (see Auer and Bernstein 2007).

The widespread availability of cochlear implants (a technology based on the electric stimulation of the auditory nerve) to palliate deafness has allowed researchers to study audiovisual integration abilities in persons who have not had prior access to (normal) acoustic input. It has been found that if an implant is introduced after infancy, deaf individuals retain a marked visual dominance during audiovisual speech perception even after years of implantation (Rouger et al. 2008). For example, Rouger et al. found that late-implanted individuals display higher incidence of visually dominant responses to the typical McGurk stimulus (that is,
reporting ‘ga’ when [ga] is presented visually and /ba/ auditorily) than normally hearing participants (whose typical fused response is ‘da’; McGurk and MacDonald 1976; and also Massaro 1998). In contrast, when the implant is introduced before 30 months of age, cochlear-implanted participants experience the McGurk effect like normally-hearing individuals (Barker and Tomblin 2004; Schorr et al. 2005). Thus findings from cochlear implanted individuals support the idea that, quite early on, the deaf brain tends to specialize in decoding speech from vision, and that this strong bias toward visual speech information seems to remain unaffected even after a dramatic improvement in auditory perception (i.e. after 6 months of implantation; see Rouger et al. 2007). However, sufficient experience at early stages of development leads to a normal (that is, not visually-biased) audiovisual integration of speech. Moreover, individuals who receive the cochlear implants seem to show a greater capacity (than healthy controls) in their capability to integrate visual information from lip-reading with the new acoustic input (see Rouger et al. 2007).

At a neural level, the primary auditory cortex of deaf individuals undergoes similar structural development (e.g. in terms of size) as in hearing individuals (see Emmorey et al. 2003; Penhune et al. 2003), thus suggesting that this area can be recruited by other modalities in the absence of any auditory input. In fact, the available evidence suggests a stronger participation of auditory association and multisensory areas (such as the superior temporal gyrus and sulcus) in deaf subjects than in normal-hearing individuals during the perception of visual speech (Lee et al. 2007) and non-speech events (Finney et al. 2001, 2003). This functional reorganization seems to occur quickly after the onset of deafness and it has been proposed that it could result from a relatively fast adaptation process based on pre-existing multisensory neural circuitry (see Lee et al. 2007). The lack of increased lateral superior temporal activity, as measured through fMRI, during visual speech perception in congenitally deaf individuals (with no prior multisensory experience; Campbell et al. 2004; MacSweeney et al. 2001, 2002) reinforces the idea that some pre-existing working circuitry specialized in integrating audiovisual speech information is needed for the visual recruitment of audiovisual multisensory areas.

Blindness, like deafness, can provide critical clues regarding the contribution of visual experience to the development of the audiovisual integration of speech. At first glance, congenitally blind individuals who have never experienced the visual correlates of speech seem perfectly able to use (understand and produce) spoken language, a fact that qualifies the importance of audiovisual integration in the development of speech perception. Some studies addressing speech production have reported slight differences between blind and sighted individuals in the pronunciation of sounds involving visually distinctive articulation, such as /p/, but these are very minor (e.g. Gólesz 1972; Mills 1987; Miner 1963; Ménard et al. 2009). Very few studies have assessed speech perception in blind individuals. In general, the results of these studies are consistent with the idea that due to compensatory mechanisms, acoustic perception in general (and as a consequence, auditory speech perception) is superior in the blind than in sighted people. In line with this idea, Lucas (1984) reported data supporting a superior ability of blind children in spotting misspellings from oral stories, and Muchnik et al. (1991) showed that the blind seem to be less affected than sighted individuals by noise in speech. If confirmed, the superiority of blind individuals in speech perception may signal a non-mandatory (or at most, minor) role of visual speech during acquisition of language. Yet, the putative superiority of speech perception in the blind remains to be conclusively demonstrated (Elstner 1955; Lux 1933, cited in Mills 1988, for conflicting findings) and, even if it were so, one would still need to ascertain whether it is speech-specific or a consequence of general auditory superiority. A detailed analysis of particular phonological contrasts in which vision may play an important developmental role will be necessary (see Ménard et al. 2009 for perception of French vowel contrasts in the blind).
9.8 Conclusions and future directions

On the one hand, we have learned a great deal about the development of multisensory speech perception over the last twenty years and, on the other, the extant findings have raised a number of new questions. By way of summary, it is now known that early emerging intersensory matching and integration abilities, some of which are based on a perceptual sensitivity to low-level intersensory relational cues (Bahrick 1983, 1988; Lewkowicz 1992a, 1992b, 1996, 2000b, 2010; Lewkowicz et al. 2010; Lewkowicz and Turkewitz 1980), can provide one basis for the initial parsing of audiovisual speech input (Kuhl and Meltzoff 1982, 1984; Patterson and Werker 1999, 2003; Pons, et al. 2009; Walton and Bower 1993). For example, early sensitivity to temporal relations could underlie sensitivity to rhythm and thus can facilitate language discrimination in infancy at the acoustic (Byers-Heinlein et al. 2010; Mehler, et al. 1988; Nazzi et al. 1998) and visual (Weikum et al. 2007) levels. It is not yet known, however, whether only low-level, domain-general mechanisms underlie the early acquisition of audiovisual speech processing skills or whether instead domain-specific mechanisms and the infant’s preference for speech also play a role. Evidence for integration of heard and seen aspects of speech in the first few months of life, albeit attenuated in its strength and in its degree of automaticity as compared to that of adults, points to a sophisticated multisensory processing of linguistic input at an early age. Converging on this conclusion, electrophysiological studies provide evidence for neural signatures of audiovisual speech integration in infancy at early stages of information processing (e.g. Bristow et al. 2008; Kushnerenko et al. 2008). Furthermore, the results of these studies suggest continuity in the development of brain mechanisms supporting audiovisual speech processing from early infancy to adulthood.

It has also become clear that, above and beyond any initial crossmodal capacities, experience plays a crucial role in shaping adult-like audiovisual speech-processing mechanisms. A finding that illustrates this tendency is that the development of audiovisual speech processing, like in the unisensory case, is non-linear. That is, rather than being purely incremental (i.e. characterized by a general broadening and improvement of initially poor perceptual abilities), recent findings (Pons et al. 2009) suggest that early audiovisual speech perception abilities are initially broad and then, with experience, narrow around the linguistic categories of the native language. This, in turn, results in the emergence of specialized perceptual systems that are selectively tuned to native perceptual inputs. This result provides evidence that the reorganization around native linguistic input not only occurs at a modality specific level, but also at a pansensory level. It will be important in the future to discern the possible inter-dependencies between unisensory and intersensory perceptual narrowing.

This development of audiovisual speech processing is also strongly driven by experience-related factors after infancy, during childhood, and into adulthood. These changes can be triggered by a number of environmental variables, such as the specific characteristics of the (native) linguistic system the speaker is exposed to (Sekiyama and Burnham 2008; Sekiyama and Tohkura 1991, 1993; Sekiyama 1997), or alterations in sensory acuity as we grow older. Indeed, when it comes to late adulthood, there seem to be some differences in audiovisual speech processing in older versus younger adults (e.g. Laurienti et al. 2006), but evidence of audiovisual integration as a compensatory mechanism for acoustic loss in old age is inconsistent at present (Cienkowski and Carney 2002; Middleweird and Plomp 1987; Sommers et al. 2005; Walden et al. 1993).

The sensory deprivation literature has provided some important findings regarding the plasticity of audiovisual speech-processing mechanisms. First, speech-reading abilities in the deaf have been found to be superior to those of hearing subjects only in limited cases (Auer and Bernstein 2007; Bernstein et al. 2000; Mohammed et al. 2005), but prior acoustic experience with speech (and hence, with audiovisual correlations) is a strong predictor for successful speech-reading in
the deaf. Second, plastic changes after cochlear implant are strongly dependent on age of implantation. Congenitally deaf individuals who have undergone cochlear implantation within the first 30 months of life achieve audiovisual integration capacities that are equivalent or even superior to those of hearing individuals (Rouger et al. 2007, 2008). When the implant occurs at an older age, a strong tendency to use visual speech information remains even after extensive experience with the newly acquired acoustic input. Visual deprivation studies are scarce and unsystematic to date, and their outcomes are mixed between mild superiority in perception (Lucas et al. 1984; Muchnik et al. 1991) and mild or no impairment in production (Mills 1987). This, does not, however, necessarily mean that visual information is not important in the perception of speech for sighted individuals who come to rely on vision in their daily life through massive visual experience.

Acknowledgements

SS-F and MC are funded by Grants PSI2010–15426 and CSD2007–00012 from Ministerio de Ciencia e Innovación and SRG2009–092 from Comissionat per a Universitats i Recerca, Generalitat de Catalunya. JN is funded by Grant PSI2009–12859 from Ministerio de Ciencia e Innovación. JFW is funded by NSERC (Natural Sciences and Engineering Research Council of Canada), SSHRC (Social Sciences and Humanities Research Council of Canada), and the McDonnell Foundation. DJL is funded by NSF grant BCS-0751888.

References


THE DEVELOPMENT OF AUDIOVISUAL SPEECH PERCEPTION


REFERENCES


REFERENCES


