Audio-Visual Speech

Some basic facts:

1) Speech recognition is more accurate when the listener can see the talker's face.
2) The speech segments that are confusable in listening are distinct visually (and vice versa).
3) Listeners integrate audio and visual information in perception.

McGurk Effect

We can create an illusion by playing a video of someone saying /ga/ while the audio plays /ba/. Listeners report hearing /da/.

If participants close their eyes, they report /ba/, so the audio signal is correctly recognized in isolation.

Theoretical Perspectives

The primitives of speech are amodal (or multi-modal?).
    This approach usually emphasizes gestures as the primitives that can be perceived based on auditory and/or visual (haptic) information. (Rosenblum)

Separate auditory and visual modules compute representational information that is then integrated (Bernstein)

The Early Integration Approach

Whether multimodal speech reflects early integration or a single multimodal stream, the argument is that the integration occurs early in processing and before phonetic perception.

Data come from a variety of studies, including Green and Miller (1985) showing that visual speaking rate influences perception of an auditory continuum. Since speaking rate extraction is early (pre-phonetic) and automatic, the argument is that audio-visual integration is early and automatic.

Early Integration Summary

Rosenblum (in Pisoni and Remez, Eds., *Handbook of Speech Perception*)

Summerfield (1987) (in Dodd and Campbell, Eds., *Hearing by Eye*)

Green (1998) (in Campbell and Dodd, Eds., *Hearing by Eye II*)

All advocate and summarize evidence for an early integration or amodal stream approach.

Separate Modules

Some results do not reconcile with amodal view and constrain the early integration view. See Bernstein (Handbook) for a summary.

One particular result, Roberts and Summerfield (1981) is particularly compelling. A McGurk display (audio /ba/, visual /ga/, perceived as /da/) is used as an adaptor in a selective adaptation study. The auditory test continuum goes from /ba/ (the same as the auditory part of the McGurk adaptor) to /da/.

Adaptation Data

As an adaptor the A/V /da/ has the same effect as the audio /ba/. If interpreted in isolation, this result can be dismissed (see Rosenblum for summary and references).

Interpreted in the context of other adaptation results, this result shows that a complex auditory coding process has taken place prior to the integration of auditory and visual information. This would imply that the amodal perspective is wrong and limits how early multimodal integration can take place.

An Aside

The adaptation data (see also Sawusch & Jusczyk, 1981; Sawusch, 1986 or Samuel, 1986 for review) also shows that intermediate representations are computed in the perception of speech. This limits the nature of any "direct perception" explanation of perception.

The system computes intervening representations. These may, however, not be available to introspection or to generate a response.