Attention, Perception, & Psychophysics
2010, 72 (1), 209-225
doi:10.3758/APP.72.1.209

The temporal distribution of information in audiovisual spoken-word identification

Alexandra Jesse and Dominic W. Massaro
University of California, Santa Cruz, California

In the present study, we examined the distribution and processing of information over time in auditory and visual speech as it is used in unimodal and bimodal word recognition. English consonant–vowel–consonant words representing all possible initial consonants were presented as auditory, visual, or audiovisual speech in a gating task. The distribution of information over time varied across and within features. Visual speech information was generally fully available early during the phoneme, whereas auditory information was still accumulated. An audiovisual benefit was therefore already found early during the phoneme. The nature of the audiovisual recognition benefit changed, however, as more of the phoneme was presented. More features benefited at short gates rather than at longer ones. Visual speech information plays, therefore, a more important role early during the phoneme rather than later. The results of the study showed the complex interplay of information across modalities and time, since this is essential in determining the time course of audiovisual spoken-word recognition.

In face-to-face communication, sensory information from the face as well as from the voice contributes to the identification of speech. The information contained in both perceptual sources is redundant but also complementary in its nature (Massaro, 1998; Sumby & Pollack, 1954; Walden, Prosek, & Worthington, 1974). For example, both hearing and seeing a speaker provide information about whether the speaker said /ba/ or /da/, but seeing the lips close for /ba/ and not for /da/ can provide more salient information than can hearing alone (Miller & Nicely, 1955; Sumby & Pollack, 1954). In addition, the co-occurrence of visual and auditory information can lead to the availability of unique multimodal cues—for example, cues to voicing (Breeuwer & Plomp, 1986; Massaro, 1998; Summerfield, 1987). In understanding audiovisual spoken words, the perceiver is therefore solving a cross-modal binding problem. Perceivers have to gather the information about a spoken word that is spread across the two modalities. Perceivers combine this information in a seemingly effortless way that leads to more robust word recognition than when they are presented with the auditory (or visual) signal alone. The size of this audiovisual recognition advantage of audiovisual over auditory-only speech is determined by the distribution of information across the auditory and visual modality—that is, by the degree to which information in the two modalities is complementary, redundant, and provides multimodal cues (see, e.g., Erber, 1974; MacLeod & Summerfield, 1987; Reisberg, McLean, & Goldfield, 1987; Sumby & Pollack, 1954). Although redundant information leads to an audiovisual recognition benefit, the benefit is larger to the degree that the two sources contain complementary rather than just redundant information (Grant & Walden, 1996; Massaro, 1998; Walden et al., 1974).

Most research in the field of audiovisual speech perception has investigated how information is processed cross-modally in order to result in an audiovisual benefit. What has been neglected in these efforts is that information about a word (and its segments) does not occur instantaneously; rather, it develops over time. Audiovisual spoken-word recognition is not only a cross-modal problem, it is also a temporal integration problem. In audiovisual word recognition, information has to be accumulated and combined over time from two modalities. Therefore, the question is not simply how visual speech information aids word recognition, but also when it does so as speech unfolds.

Understanding the distribution of audiovisual information over time is critical, because it must influence the time course of audiovisual spoken-word recognition. Research on auditory word recognition has shown that incoming speech information has an immediate and continuous influence on lexical access and competition (see, e.g., Davis, Marslen-Wilson, & Gaskell, 2002; Tyler, 1984; Zwitserlood, 1989). Words are activated to the degree that they match the incoming signal. This degree of overlap is continuously updated over time on the basis of incoming information. For example, with the availability of disambiguating information between *candy* and *candle*, listeners start to look more often at a picture of the target word *candy* than of the competitor word *candle* (Allopenna, Magnuson, & Tanenhaus, 1998). However, word competition is modulated over time by the similarity of word representations with the input. Rhyme competitors, such as *sandy*, that become more similar later on to the target *candy* will have

A. Jesse, alexandra.jesse@mpi.nl

a later influence on word recognition than will onset competitors, such as *candle* (Magnuson, Dixon, Tanenhaus, & Aslin, 2007). That is, the time course of word recognition and the time course of lexical competition are governed by the temporal distribution of information in the signal, and its resulting change in similarity to the word representations is stored in our mental lexicon.

Consequently, the distribution of information over time in the auditory speech signal has been extensively documented, since it is a prerequisite to understanding auditory word recognition (see, e.g., Smits, Warner, McQueen, & Cutler, 2003). However, little is known about the temporal distribution of information in visual speech and its relationship to the availability of auditory information over time. In audiovisual word recognition, the auditory as well as the visual lexical structure have an influence on the recognition of a word (Auer, 2002; Brancazio, 2004; Mattys, Bernstein, & Auer, 2002). It is therefore to be expected that in audiovisual word recognition, the degree of support for lexical candidates is continuously updated on the basis of auditory as well as visual speech information to resolve lexical competition. To understand the dynamics of audiovisual spoken-word recognition, it is necessary to investigate the distribution of information over time in auditory and visual speech. In the present study, we provided this crucial basis to further our understanding of audiovisual spoken-word recognition by tracking the distribution of audiovisual speech information over points in time, since it is used unimodally and in audiovisual word recognition.

The Distribution of Auditory and Visual Information Over Time

The acoustic signal is a direct consequence of the articulators' configurations and movements; changes in the articulators ought to be closely linked in time to changes in the acoustic signal. However, the informativeness of these changes in the two signals is not necessarily closely linked. First of all, not all articulatory changes are visible. For example, the vibration of the vocal cords leads to cues about voicing in the acoustic signal, but only to weak voicing cues in the optical signal (Yehia, Rubin, & Vatikiotis-Bateson, 1998). Second, visual information can precede auditory information. Information about vowel identity can be already available about 160 msec before the acoustic onset of the vowel (Cathiard, Lallouache, Mohamadi, & Abry, 1995). Likewise, seeing the lips close to prepare a release in the production of a bilabial plosive, such as /b/, is accompanied by silence. Seeing the lips closing, however, is sufficient for robust recognition of the labial place of articulation (Smeele, 1994). That is, the visual signal provides this information before the auditory signal does. Even though information from the two modalities arrives at different times, their integration is relatively robust to these cross-modal asynchronies (R. Campbell & Dodd, 1980; Massaro, 1998; Massaro & Cohen, 1993; Massaro, Cohen, & Smeele, 1996; Munhall, Gribble, Sacco, & Ward, 1996; van Wassenhove, 2004).

Further evidence that the information in visual and auditory speech does not unfold at the same rate was found in a gated version of a McGurk study (McGurk & MacDonald, 1976; Munhall & Tohkura, 1998). In the McGurk effect, an audiovisual vowel–consonant–vowel stimuli consisting of a visual /ægæ/ and an auditory /æbæ/ commonly leads to an /ædæ/ percept since, overall, a /d/ best matches the information provided by both modalities (McGurk & MacDonald, 1976). In the gated version, Munhall and Tohkura presented increasingly longer segments of the visual stimulus /ægæ/ in combination with the complete auditory stimulus /æbæ/. The number of McGurk percepts (/ædæ/) increased linearly with an increase in visual information. However, for the situation in which the auditory but not the visual signal was gated, the number of McGurk percepts increased suddenly only at the gate that included the release burst of the plosive; then, it stayed at this level for the remaining gates. Munhall and Tohkura concluded that the information in visual and auditory speech does not seem to unfold at the same rate. Visual information is distributed linearly, whereas the auditory signal varies in its informativeness rapidly and nonlinearly. However, the validity of this conclusion is limited by the fact that only plosive sounds were used as stimuli materials in a single context. Information for other phonemes can be more evenly distributed in the auditory signal over time (Smits, 2000; Smits et al., 2003).

The results of these studies suggest that information in the two modalities becomes available at different points in time and, furthermore, that the information value of the two modalities changes differently over time. Consequently, this means for audiovisual word recognition that the benefit obtained from the addition of visual speech on the recognition of words should be modulated over time, depending on the differential temporal distribution of information in the two modalities. That is, the time course of the audiovisual benefit is modulated by the distribution of auditory and visual information over time. Audiovisual spoken words will therefore differ not only in the degree to which visual speech contributes to their recognition, but also in when visual speech provides a benefit.

To understand these internal dynamics of audiovisual spoken-word recognition, in the present study, we systematically tracked over time the availability of information across and within auditory, visual, and audiovisual speech for the recognition of (American) English words. An exhaustive set of 66 consonant–vowel–consonant (CVC) words representing all possible initial consonants in English was selected. Eight different vowel contexts were tested to enhance the generalizability of the results. An audiovisual extension of the gating task (De la Vaux & Massaro, 2004; Munhall & Tohkura, 1998; Seitz & Grant, 1999; Smeele, 1994) was used to systematically manipulate the availability of information. Participants were presented with auditory-only, visual-only, and audiovisual word fragments of varying length, all starting from the onset of the word. The gating task is an appropriate and widely used task to establish the time course of information becoming available in auditory speech (see, e.g., Grimm, 1966; Grosjean, 1980; Kiefte, 2003; Öhman, 1966; Pols & Schouten, 1978; Smits, 2000; Smits et al., 2003; for a similar argument; see Grosjean, 1996).

The goal of the present gating study was to establish the time course of auditory, visual, and audiovisual information in spoken-word recognition. Although a few recent studies have examined the temporal distribution of information in an audiovisual gating task, none of these studies allowed for a systematic analysis of the availability of information over time, since either only a few word items were used (Munhall & Tohkura, 1998; Seitz & Grant, 1999; Smeele, 1994), or different words (and different consonants) were presented at each gate (De la Vaux & Massaro, 2004). Thus, even though the studies showed a general audiovisual benefit for gated words, these studies were not informative about the distribution of audiovisual information over time, nor about how this distribution determines the time course of the audiovisual recognition benefit of words. By testing an exhaustive set of words representing all possible initial consonants in English as stimulus materials, the present study made it possible for the first time to examine when phoneme and featural information become available in the speech signals and to trace this information not only across modalities, but also over gates. The study allowed for a detailed investigation of the changes in featural contributions to the audiovisual benefit while speech unfolds. In addition, the study provides a database for the development and testing of quantitative models of the time course of audiovisual spoken-word recognition. The collected data supply models with estimates of segmental information availability in the modalities over time, needed by the models to capture the time course of audiovisual word recognition. A successful example of this modeling strategy can be found for auditory word recognition in the literature (Norris & McQueen, 2008). The recordings also add to the limited number of audiovisual speech corpora and allow for parametric analyses of audiovisual speech.

To attain these multiple goals of the present study, we used synthetic speech to create testing materials. Synthetic speech fulfills the simultaneous need for highly controllable and reproducible stimuli, as well as for a high and flexible temporal resolution of the time course. The stimulus materials can be exactly reproduced or systematically manipulated for comparisons in future studies (e.g., to test the robustness of the results across multiple viewing positions or contexts). An unlimited number of stimulus sets can be produced. This provides a substantial benefit over most more limited audiovisual speech corpora, since it allows for parametric analyses without exhausting the stimulus set. Furthermore, synthetic speech, through its high and flexible temporal resolution, enables fine-grained analyses of the temporal distribution within and across the speech signals. To provide fine-grained time course data, a gating study needs to tap into the speech stream very early and often so that it can capture the accumulation of speech information (Kiefte, 2003; Smits et al., 2003; Stevens & Blumstein, 1978; Tekieli & Cullinan, 1979). The temporal flexibility also allows producing and reproducing stimuli at any gate duration. This is also critical to ensure an equal number of time slices (gates) for each stimulus, which is a necessary prerequisite for subsequent quantitative modeling (Norris & McQueen, 2008). Synthetic speech was therefore necessary to address the research questions posed and to increase the data's importance for future research.

METHOD

Participants

A total of 130 monolingual native English speakers who reported no hearing, vision, or language deficits participated for course credit or pay. All of the participants were undergraduate students at the University of California, Santa Cruz, and their age range was 18–30 years old (average age = 20).

Materials

Sixty-six English CVC words were selected that consisted of three tokens of each of the 22 possible initial consonants in English (see the Appendix). The average written-word frequency (Kučera & Francis, 1967) of these words was $M = 275$, $SD = 1{,}336$ (without the outlier THAT: $M = 114$, $SD = 311$). Their Switchboard spoken frequency was $M = 23$, $SD = 134$ (Greenberg, 1999). On average, these words had 13.35 phonotactic neighbors ($SD = 6.26$), with 5.9 of them being on average higher frequency neighbors ($SD = 4.32$). Each token within an initial consonant set was paired with a different subsequent vowel (/æ/, /ɛ/, /ɑ/, /ɔ/, /i/, /ɪ/, /u/, /ʊ/). Vowels within a consonant set were visually distinguishable; that is, they all belonged to different viseme classes ({æ,ɛ}, {ɑ,ɔ}, {i,ɪ}, {u}, and {ʊ}; see Massaro, 1998, p. 395). Exceptions had to be made for words starting with /θ/, /v/, /z/, and /ð/. Overall, each vowel occurred approximately equally often in the word list. The nature of the final consonant in these CVC words was not controlled. No rime was repeated, with the exception of one repetition of /il/, /æt/, /ʊd/, and /ʊk/. All of the words within each initial phoneme set had the same spelling for the initial phoneme, with the exception of CZAR. As was pointed out to the participants, the response button for CZAR was grouped with those for words starting with "Z." All of the items were preceded by the word "a" (/ʌ/), with an average duration of 92.67 msec. This was done to ensure that the initial gates of plosives and /h/ could be used (Smits et al., 2003). This preceding context was held constant, even though it sometimes provided an ungrammatical utterance (e.g., "a cash," "a that"). Participants were instructed to ignore the overall grammaticality of the stimuli as well as the context word itself, and were specifically reminded to do so in the visual condition.

Recording and Gating

All of the items were recorded with the BAPI software that controls the virtual talker of the study, Baldi (Massaro, 1998). Baldi was driven by a synthetic male voice of American English (NeoSpeech SAP15, Paul, M, 16000/16), which was based on the NeoSpeech text-to-speech system (www.neospeech.com/product/data/VoiceText.pdf). The facial animation system uses phonemes as basic units of synthesis (see Massaro, 1998, chap. 12, for an overview). Each phoneme is represented in the model as a set of target values of facial control parameters (e.g., jaw rotation, mouth width). Phonemes are concatenated by the model following rules for coarticulation, where the relative dominance of the parameters for each speech segment is varied as a function of context (Löfqvist, 1990; Saltzman & Munhall, 1989). Temporal dominance functions control for each parameter of a segment how much weight the parameter's target value carries against those of all preceding and upcoming segments, and therefore determines how target values are blended over time. Measurements of natural productions are used to set the target values and to specify their change over time relative to the auditory signal. This PSL–UCSC coarticulation algorithm has been successfully used in American English as well as in several other languages (see Massaro, 1998).

Videos for each token at each modality and gate combination were recorded separately as .avi files, with a rate of 60 frames/sec for the

video and of 16 kHz for the auditory channel. Videos were displayed centered on a black background in a 208 × 300 pixel window. For auditory-only recordings, a black bar covered the video completely, so that the screen appeared to be completely black. All of the videos started with 200 msec of silence that was accompanied with the display of the speaker in a resting position for tokens in conditions with visual input. The talker showed no eyebrow or head movements and did not blink. All videos ended in a black screen to gain better control over the duration of the stimulus presentation.

Stimuli were directly produced as gated tokens during the recording. This was done separately, but with the same gate durations for the three modality conditions. Gates were created as thirds of the respective duration of the initial consonant (M = 35 msec, SD = 12 msec, range = 12–64 msec) and the vowel (M = 57 msec, SD = 20 msec, range = 20–103 msec). At the third gate, the presentation of the initial consonant was complete; at the sixth gate, the presentation of the consonant–vowel (CV) was complete. The proportional method of gating was preferred over fixed gating (e.g., every 40 msec), in order to ensure the same number of gates for each word. Phoneme boundaries were determined by the BAPI software and verified by a phonetically trained researcher following Stevens's (2000) description of acoustic phonetics. The onset of stops and affricates was defined as the beginning of the stop closure or the beginning of prevoicing. This meant that any preparatory coarticulatory movements—for example, the preparation of the closure of the lips during the preceding vowel—were included in the first gate. The end of stops was defined as the offset of their bursts. The end of affricates was defined as the end of frication. Fricatives were determined on the basis of the onset and offset of their frication noise. Nasals were defined by their characteristic drastic reduction of energy in the spectrum. For approximants, the boundaries were set to the middle of the transition periods with the surrounding vowels. Since liquids followed a vowel in the present study, the same criteria as those for approximants were applied.

To avoid a bias to report the abrupt offset of gated stimuli as labials or plosives (Pols & Schouten, 1978), the auditory signal was always ramped linearly from 100% to 0% during the last 5 msec before the offset of a gate. The linear ramp was applied during the recording to ensure that the auditory and visual signals did not need to be realigned.

Apparatus

Participants were tested individually in one of four sound-treated testing rooms with identical equipment. Auditory stimuli were presented over Plantronics Audio 90 headsets at a similar comfortable hearing level. The experiment was controlled in each room by the Rapid Application Developer in the CSLU Toolkit on PCs with NVIDIA GeForce3 TI500 64MB video cards and Creative SB Live! sound cards.

Procedure

The task of the participants was to identify the word that best matched the word onset they perceived. Participants were told that on some trials, they could see and hear a speaker talk, whereas on others, they would only see or only hear the speaker. It was stressed that at all times they were to watch the computer screen, which was approximately 50 cm in front of them. Participants were familiarized with the arrangement of the response buttons. The experimenter emphasized that participants should remember to consider all possible word alternatives before responding, and that they were to ignore the preceding context word "a." During a practice block, participants were then familiarized with all target words as complete audiovisual stimuli presented in randomized order. In all practice and experimental trials, a black screen was initially presented for 1,400 msec, followed by a fixation cross for 600 msec. The video was then displayed horizontally centered and 150 pixels below the top of the screen. The video player had no surrounding frames. The mouse cursor was not visible on top of the video. All of the video presentations ended in a black screen and with the simultaneous presentation of 66 response buttons corresponding to the possible word choices on the lower portion of the screen. Buttons were ordered in a matrix by initial consonantal phoneme. The time to respond was unlimited. Once a response was given, the screen blackened again, and the next trial started.

The main experiment consisted of three equal-sized blocks of trials. Blocks were balanced on modality conditions, gates, and words, so that each word was presented six times within each block: twice in each modality condition and once at each gate. Each participant was tested on all 66 words under all conditions once, and therefore responded to a total of 1,188 trials. The order of trials within each block was randomized. The order of blocks was counterbalanced across participants. Participants took up to 5-min-long breaks between the blocks of the approximately 3-h-long experiment.

RESULTS

Nineteen participants did not complete or comply with the experiment for various reasons. Of the remaining participants, three individual responses were set to missing values due to equipment failure. This resulted in a total of 111 participants providing 131,865 data points.

The Time Course of Phoneme and Viseme Recognition

Figure 1 shows the percentage of correct consonant recognition in the three modality conditions over gates. ANOVAs on the average percentage-correct identification of initial consonants, with modality (visual only, auditory only, audiovisual) and gate condition (six gate levels) as within-subjects factors, were conducted. Generalized η^2 measures are reported as indicators of effect size (Bakeman, 2005; Olejnik & Algina, 2003).

ANOVAs on the percentage of correctly recognized initial-consonant phonemes showed a significant main effect of modality [$F(2,220) = 26{,}782.75$, $p < .001$, $\eta^2_G = .97$] and of gate [$F(5,550) = 998.34$, $p < .001$, $\eta^2_G = .53$], as well as a significant interaction between these two factors [$F(10,1100) = 171.62$, $p < .001$, $\eta^2_G = .25$]. Performance thus varied across modality conditions. Recognition performance also changed over gates, but did so differently depending on the speech modality. To further assess the time course of information processing within each of the three speech-modality conditions, planned comparisons with a Bonferroni-corrected α level compared performance at adjacent gates in each modality (see Table 1 for details). These analyses showed that phoneme recognition improved over all gates in both the auditory-only and audiovisual conditions. Phoneme recognition performance in the visual condition did not substantially change. Correct visual phoneme recognition increased only between Gates 3 and 4. That is, visual information about the following vowel was used to successfully improve consonant phoneme recognition. The three word tokens of each given consonant class contained subsequent vowels from different viseme classes. Perceivers could have used visual speech information also to benefit from the fact that only a certain set of CV combinations were given as response alternatives.

Performance in the visual speech condition was generally expected to be lower than what was previously found in nongating studies, because gating severely degrades

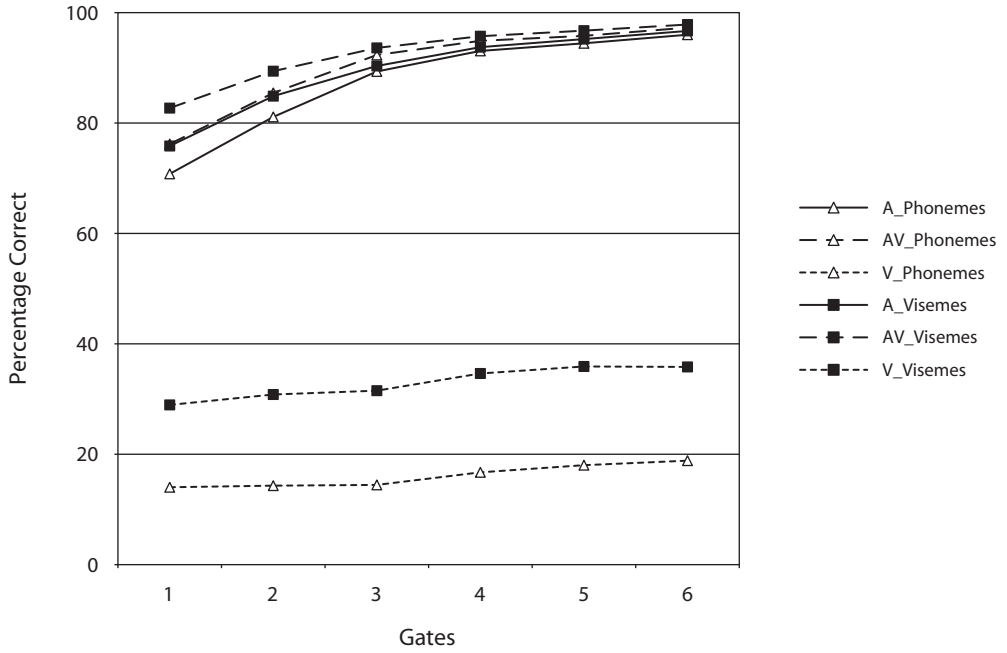

Figure 1. The percentage correct for initial consonant phoneme and viseme recognition in auditory (A), visual (V), and audiovisual (AV) speech over gates.

the visual signal by limiting the availability of information. Previous studies also commonly tested only a limited set of consonants that were chosen to be visually easy to distinguish—that is, taken from different viseme classes. Phonemes within a viseme class are visually not or only somewhat distinguishable from one another (Fisher, 1968). Phonemes from different viseme classes are highly visually distinguishable from one another. Correct consonant identification was also assessed in terms of correct recognition of viseme groups (see Table 6 for viseme grouping; Massaro, 1998, p. 395). A reanalysis of the present data pooled over consonant phonemes that belonged to the same viseme group showed for visual speech a recognition rate for visemes of 29% correct at Gate 1 and of 36% correct at Gate 6 (see Figure 1). This suggests that the quality of the visual speech is good and comparable to that in previous nongating studies using only one phoneme from each viseme class as stimuli. Note that, relative to previous studies, poorer performance was still expected to be found in this analysis, because we also included visemes with generally low visibility, such as {h} and {j}, averaging below 5% in the present study.

Statistical analyses on the average percentage of correctly recognized initial visemes showed the same general result pattern as that found for the phoneme recognition analyses. ANOVAs on the percentage of correctly recognized initial visemes showed a significant main effect of modality [$F(2,220) = 4{,}847.07$, $p < .001$, $\eta_G^2 = .93$] and of gate [$F(5,550) = 594.74$, $p < .001$, $\eta_G^2 = .30$], as well as a significant interaction between these two factors [$F(10,1100) = 59.08$, $p < .001$, $\eta_G^2 = .06$]. Performance varied not only across modality conditions, but also across gates, and did so differently depending on the modality condition. As had been the case for phoneme recognition, planned comparisons showed that the correct recognition of viseme class improved over all gates in the auditory-only and audiovisual conditions (see Table 1). Visual viseme recognition improved significantly only between

Table 1
Pairwise Comparisons of the Percentage of Correct Consonant Phoneme and Viseme Recognition Across Gates for Each Modality Condition

		Gates									
	Modality Condition	1 vs. 2		2 vs. 3		3 vs. 4		4 vs. 5		5 vs. 6	
		t	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
Phonemes	Auditory only	21.25	<.001	17.92	<.001	11.33	<.001	5.56	<.001	5.67	<.001
	Audiovisual	19.99	<.001	16.18	<.001	7.12	<.001	3.71	<.001	6.53	<.001
	Visual only	0.57	.57	0.29	.78	4.12	<.001	2.17	.03	6.53	.14
Visemes	Auditory only	18.19	<.001	13.17	<.001	11.19	<.001	6.02	<.001	5.58	<.001
	Audiovisual	15.74	<.001	11.98	<.001	6.84	<.001	4.28	<.001	5.02	<.001
	Visual only	2.76	.007	1.17	.24	4.88	<.001	2.13	.036	−0.14	.89

Note—The degree of freedom for all *t* tests was 110.

Table 2
Feature Classification Scheme for Consonants

Features	b	p	m	f	v	θ	ð	t	d	n	k	g	s	z	ʃ	tʃ	dʒ	h	j	l	r	w
Voicing	+	−	+	−	+	−	+	−	+	+	−	+	−	+	−	−	+	−	+	+	+	+
Nasality	−	−	+	−	−	−	−	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−
Place	0	0	0	0	0	1	1	1	1	1	2	2	1	1	2	2	2	2	2	1	1	2
Frication	−	−	−	+	+	+	+	−	−	−	−	−	+	+	+	+	+	+	−	−	−	−
Duration	−	−	−	−	−	−	−	−	−	−	−	−	+	+	+	+	+	−	+	+	+	+
Rounding	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−	+	+
Continuant	−	−	−	+	+	+	+	−	−	−	−	−	+	+	+	−	−	+	+	+	+	+

Note—"+" indicates voiced, nasalized, and fricative of long duration, rounded, and continuant, respectively. Place of articulation is coded as 0 = front place (bilabial and labiodental), 1 = mid place (dental and alveolar), and 2 = back place (postalveolar, palatal, velar, and glottal).

Gates 1 and 2. Visual discrimination among visemes thus improved early on, whereas the visual discrimination among all phonemes improved only with some additional (vowel) information about the word.

The Audiovisual Phoneme Recognition Benefit and Its Time Course

To examine the audiovisual benefit, an overall ANOVA with the factor gate (six levels) was conducted on a relative audiovisual benefit measure. This relative audiovisual benefit was calculated at each gate as the amount of improvement between the audiovisual and auditory conditions set in relation to the overall possible amount of improvement [(audiovisual–auditory)/(100%–auditory); Sumby & Pollack, 1954]. This audiovisual benefit measure takes into account that a benefit is more difficult to observe when performance in the auditory-only condition approaches its upper limit. These analyses were conducted on correct phoneme recognition. Overall, the audiovisual benefit did not vary across gates [$F(3.6, 396.04) = 1.476$, $p < .21$, $\eta_G^2 = .01$; given a violation of the sphericity assumption, Greenhouse–Geisser corrected degrees of freedom are reported here]. Planned comparisons with a Bonferroni-corrected α level tested for the existence of an audiovisual recognition benefit at each gate. These one-sample t tests showed that an audiovisual benefit was observed at each gate [Gate 1, $t(110) = 11.35$, $p < .001$; Gate 2, $t(110) = 9.33$, $p < .001$; Gate 3, $t(110) = 8.96$, $p < .001$; Gate 4, $t(110) = 5.63$, $p < .001$; Gate 5, $t(110) = 3.60$, $p < .001$; Gate 6, $t(110) = 4.89$, $p < .001$]. A reliable audiovisual phoneme recognition benefit was therefore observed at each gate, and this benefit did not vary in size over time.

The Distribution of Featural Information

As was shown previously, seeing and hearing the speaker helped phoneme recognition to about the same degree at each gate. We next examined how information about linguistic features is distributed across modality conditions as well as over the speech signal within a modality condition. A particular focus of the present study was to assess how these featural distributions across modality conditions and gates alter the audiovisual benefit while speech unfolds.

To examine the nature of information that becomes available for word recognition in the three modality conditions and its contribution to the audiovisual benefit over time, the percentage of transmitted information (%TI; Shannon, 1948) for a set of linguistic features was calculated. This type of feature analysis assesses the transmission of linguistic features (e.g., duration) and not the transmission of low-level acoustic and optical properties (e.g., length). It is thus not concerned with what property of the signal transmits information about a phoneme's linguistic feature. %TI is a bias-free measure that describes the relationship between stimuli and responses. Guessing performance leads to a %TI equal to 0, irrespective of the number of response alternatives. The more information transmitted in the signal that aids to discriminate between the possible responses, the closer %TI is to 100%. Because performance approached ceiling level around the fourth gate, only confusions at the first three gates were considered. These gates fall within what had been determined to be the boundaries of the initial consonants. Therefore, a set of features related to consonants was analyzed on the basis of consonant confusion matrices.

In order to uniquely define each consonant by its feature specification (see Table 2), the linguistic feature set consisted of the features voicing, nasality, place of articulation, frication, and duration, as defined by Miller and Nicely (1955). The features rounding and continuant were added (Chomsky & Halle, 1968). Rounding is of interest, since this feature should be visible during speech production (Benguerel & Pichora-Fuller, 1982; Lisker & Rossi, 1992; Robert-Ribes, Schwartz, Lallouache, & Escudier, 1998; Traunmüller & Öhrström, 2007). The feature continuant reflects whether a phoneme is produced with or without complete blocking of the airflow and aids the further coding of manner of articulation. Definitions for /tʃ/, /h/, /j/, /l/, /r/, and /w/ were added.

%TI was calculated separately for each participant for each featural confusion matrix in each modality condition. For each feature analysis, the consonant confusion matrices of each participant were converted into smaller matrices in which rows and columns for phonemes that shared a feature were grouped together (Miller & Nicely, 1955). For place of articulation, the data in each confusion matrix were regrouped by whether a phoneme had a front, mid, or back place of articulation, resulting in a 3×3 matrix. For all other features, the data were regrouped in 2×2 matrices (i.e., phonemes either had the feature or not).

Figure 2 shows the transmission values for each feature in each of the three modality conditions over gates. A series of ANOVAs on the percentage of transmitted information for each feature, with modality and gate as

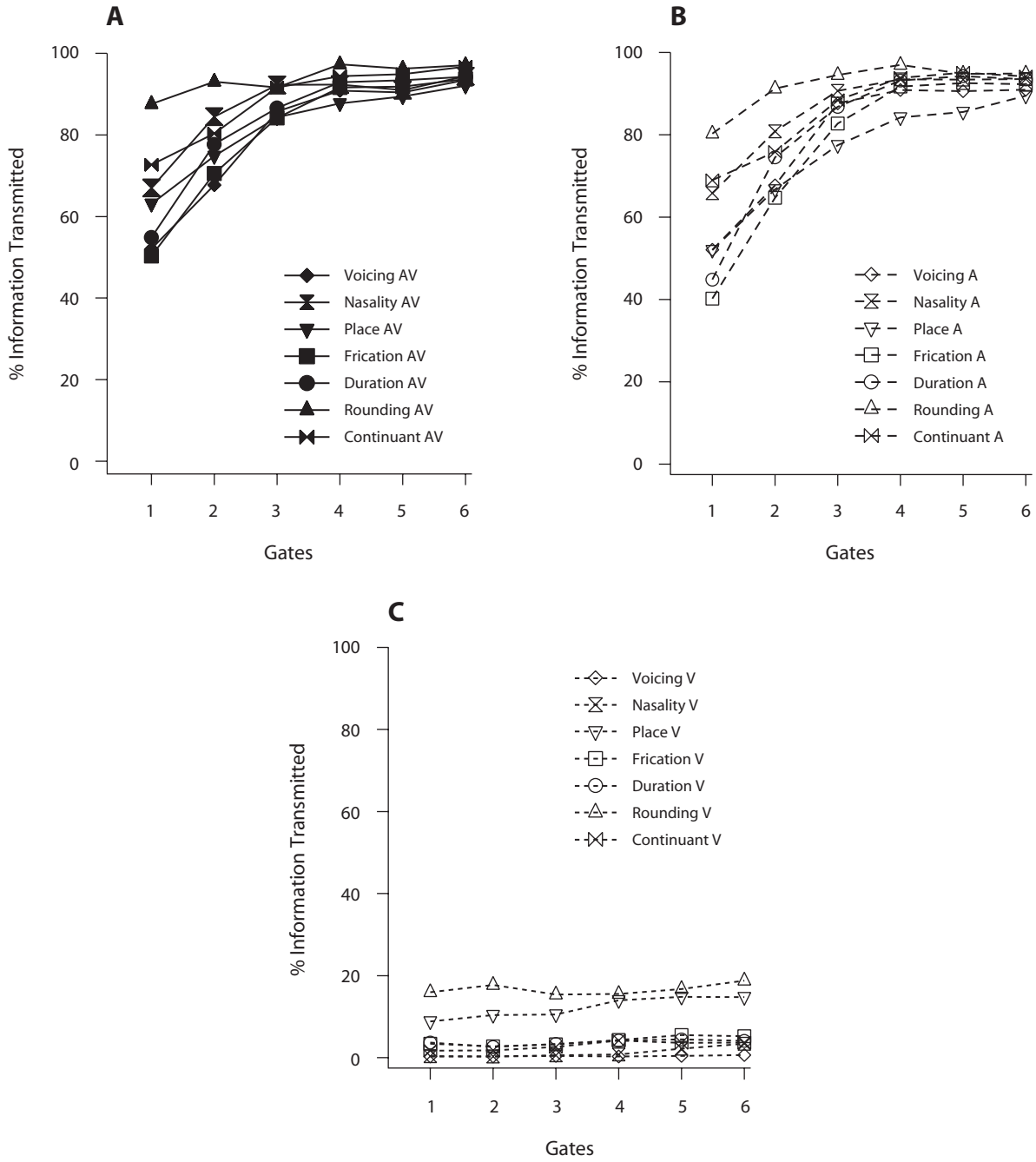

Figure 2. The percentage of information transmitted for the features voicing, nasality, frication, duration, place of articulation, rounding, and continuant, based on consonant confusion data in (A) audiovisual, (B) auditory, and (C) visual speech over gates.

within-subjects factors, showed significant main effects of modality and gate as well as their significant interaction for all features. Table 3 summarizes these results. Figure 2 shows a higher percentage of transmission of all linguistic features in the auditory than in the visual signal from the first gate on. The figure suggests that at the first gate, the auditory input was the most informative about the features rounding, continuant, and nasality. The poorer transmitted features were frication, duration, voicing, and place of articulation. The visual signal mostly transmitted information about rounding and place of articulation. For all other features, the %TI for the visual signal stayed always below about 5%.

Planned comparisons assessed the change of the percentage of transmitted information for each feature over the first three gates within each modality (see Table 4). These comparisons showed that the transmission values for nearly all of the features increased over gates for audi-

Table 3
Results Overview of ANOVAs Conducted on the Percentage of Featural Information Transmitted for Each Feature

	Modality			Gates			Modality × Gates		
Features	$F(2,220)$	p	η^2_G	$F(2,220)$	p	η^2_G	$F(4,440)$	p	η^2_G
Voicing	4,070.41	<.001	.90	451.31	<.001	.42	172.89	<.001	.26
Nasality	3,427.84	<.001	.89	128.34	<.001	.17	54.76	<.001	.10
Place	2,404.87	<.001	.84	405.84	<.001	.22	82.89	<.001	.09
Place front	444.02	<.001	.56	38.69	<.001	.02	3.23	<.01	.004
Place mid	1,727.42	<.001	.82	395.18	<.001	.26	101.75	<.001	.12
Place back	5,328.90	<.001	.91	543.83	<.001	.41	173.78	<.001	.25
Frication	3,067.59	<.001	.87	605.13	<.001	.44	232.14	<.001	.30
Duration	4,248.74	<.001	.89	616.63	<.001	.46	232.50	<.001	.32
Rounding	1,264.95	<.001	.79	22.26	<.001	.02	7.53	<.001	.02
Continuant	4,366.40	<.001	.91	180.92	<.001	.15	39.31	<.001	.07

torily only and audiovisually presented speech. Only the amount of transmitted information for rounding increased solely from Gates 1 and 2, but it remained the same between Gate 2 and Gate 3. The pattern for the visual signal, however, differed. For nearly all features, the %TI approached an asymptotic value early on. This contrasts with the auditory and the audiovisual cases, in which the informativeness increased over additional gates. An exception is the place of articulation information, however. The transmitted information for place of articulation increased between Gates 1 and 2 before remaining at an asymptotic level. In summary, visual and auditory speech differed in their featural informativeness, but this difference varied across the speech signal.

In the present study, it seems that there was somewhat more information for rounding than for place of articulation in visual speech. The latter is usually found as the best transmitted feature in visual speech, although information about rounding is also accessible (Benguerel & Pichora-Fuller, 1982). However, previous studies often did not include rounding as a feature. Furthermore, the current %TI measure for place indicates how much information there is to distinguish between front, mid, and back places of articulation. The analysis simply measured how much all places of articulation are discriminable from one another. The visual signal, however, should mostly provide information to discriminate whether a consonant was produced in the front of the oral cavity or not. The mid and back places of articulation should be more confusable. This confusability between the two places consequently lowered the %TI value for place of articulation overall. An additional analysis examined, therefore, how much information was available to discriminate one place from all of the other places. The confusion data were pooled for each participant according to whether a consonant has, for example, a front place of articulation. Data for consonants with a mid or back place of articulation were pooled into the category "not front place" for this analysis.

These additional place analyses showed that there was approximately the same amount of information available in the visual signal to recognize front place of articulation (over gates: 16%, 18%, and 19%) as there was to recognize rounding (16%, 18%, and 15%). %TI levels for mid and back place information in the visual signal remained below 5%. In comparison, the auditory signal mostly contained place information that helps perceivers to distinguish the back place of articulation (54%, 77%, and 92%) from all other places. Mid place information in the auditory signal ranged from 36% to 62%, and front place information ranged from 54% to 63% over the first three gates. In the combined audiovisual signal, the front and back place of articulation information was mostly transmitted. Transmission values for front place ranged from 68% to 77%, and for back place, they ranged from

Table 4
Pairwise Comparisons of the Percentage of Information Transmitted of Consonantal Features Across the First Three Gates for Each Modality Condition

	Auditory Only				Audiovisual				Visual Only			
	Gates 1 vs. 2		Gates 2 vs. 3		Gates 1 vs. 2		Gates 2 vs. 3		Gates 1 vs. 2		Gates 2 vs. 3	
Features	t	p	t	p	t	p	t	p	t	p	t	p
Voicing	12.49	<.001	13.80	<.001	11.34	<.001	11.77	<.001	−0.28	.78	1.41	.16
Nasality	8.22	<.001	5.41	<.001	8.47	<.001	4.65	<.001	−1.77	.08	0.37	.71
Place	14.28	<.001	10.04	<.001	12.92	<.001	9.13	<.001	3.01	.003	0.02	.98
Place front	3.00	.003	2.97	.004	2.01	.05	4.20	<.001	3.59	<.001	−0.61	.54
Place mid	13.54	<.001	8.63	<.001	12.40	<.001	8.80	<.001	3.72	<.001	−0.43	.67
Place back	16.44	<.001	11.11	<.001	15.12	<.001	9.10	<.001	0.60	.55	0.87	.39
Frication	16.69	<.001	12.96	<.001	16.60	<.001	10.75	<.001	−1.31	.19	−0.04	.97
Duration	20.94	<.001	9.47	<.001	15.81	<.001	7.53	<.001	−1.10	.27	0.27	.79
Rounding	6.67	<.001	1.63	.11	3.35	<.001	−0.74	.46	1.04	.30	−1.24	.22
Continuant	4.65	<.001	9.71	<.001	4.79	<.001	10	<.001	1.48	.14	0.79	.43

Note—The degree of freedom for all t tests was 110.

Table 5
Evaluations of the Relative Audiovisual Benefit of Transmitted Information of Consonantal Features for the First Three Gates, As Well As Pairwise Comparisons of the Size of This Benefit Across Gates

	Gates									
	1		2		3		1 vs. 2		2 vs. 3	
Features	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
Voicing	−2.03	.05	−1.07	.29	1.47	.14	0.24	.81	1.80	.08
Nasality	2.10	.04	9.79	<.001	10.23	<.001	4.30	<.001	0.66	.51
Place	11.61	<.001	8.92	<.001	9.45	<.001	1.35	.18	2.12	.04
Place front	9.80	<.001	8.21	<.001	10.82	<.001	0.11	.91	2.11	.04
Place mid	11.79	<.001	8.96	<.001	10.86	<.001	1.95	.05	3.51	.001
Place back	4.51	<.001	4.00	<.001	8.15	<.001	1.00	.32	2.75	.007
Frication	4.39	<.001	2.19	.03	3.31	.001	−0.55	.59	1.49	.14
Duration	6.03	<.001	0.99	.32	2.70	.008	−1.58	.12	1.34	.18
Rounding	5.82	<.001	8.90	<.001	7.83	<.001	1.00	.32	−0.13	.89
Continuant	0.69	.49	1.43	.16	5.55	<.001	0.51	.61	2.58	.01

Note—Positive *t* values indicate an audiovisual benefit and an increase of the audiovisual benefit across gates. The degree of freedom for all *t* tests was 110.

62% to 92% over the first three gates. In comparison, information for mid place ranged from 50% to 75%. Planned comparisons (see Table 4) suggested further that the distribution of this subfeatural information over gates varied across modality conditions. Auditory and audiovisual mid and back place information increased over the first three gates, whereas front place information did not. This result complements the visual signal, in which visual front and mid place information increased over the first two gates. Visual and auditory speech thus seem to differ not only in their informativeness across features, but also in what information they provide for a given feature and when this information becomes available during the speech signal.

The Featural Contribution to the Audiovisual Phoneme Recognition Benefit and Its Time Course

To examine the featural contribution to the audiovisual recognition benefit and its time course over the speech segment, the relative audiovisual benefit, based on the percentage of transmitted featural information, was calculated for each linguistic feature. Table 5 shows results from planned comparisons evaluating the audiovisual benefit for each feature at each gate and the change across gates. Audiovisual benefits for place of articulation (overall M = 25%) and rounding (M = 23%) were found at each gate. Duration (M = 16%) and frication (M = 17%) showed an audiovisual benefit at Gates 1 and 3, but not at Gate 2. Nasality (M = 9%) showed an audiovisual benefit only at the two later gates, and the feature continuant (M = 17%) only at Gate 3. There was never sufficient visual information about voicing to contribute substantially to an audiovisual benefit. Note that the size of all of these featural audiovisual benefits rarely varied across gates. One exception is nasality, which showed an increase in audiovisual benefit between the first two gates. The audiovisual benefit for the feature continuant and place tended to increase somewhat late between the second and third gates.

As is shown in the above analyses, auditory and visual speech provide different place of articulation information and do so also to different degrees over time. Auditory speech mainly transmitted back place information; visual speech contains mainly front place information. The audiovisual signal transmitted mainly both back and front place information. Analyses on the audiovisual place benefit showed that an audiovisual benefit for all three places of articulation can be found at each gate (front place, M = 33%; mid place, M = 25%; back place, M = 17%). That is, complementarity as well as redundancy of the featural information transmitted by the two speech signals led to an audiovisual benefit. The audiovisual recognition benefit for each place of articulation was readily available in substantial size for all three place features (front place, $M_{\text{Gate 1}}$ = 35%, $M_{\text{Gate 2}}$ = 35%, $M_{\text{Gate 3}}$ = 45%; mid place, $M_{\text{Gate 1}}$ = 25%, $M_{\text{Gate 2}}$ = 31%, $M_{\text{Gate 3}}$ = 44%; back place, $M_{\text{Gate 1}}$ = 17%, $M_{\text{Gate 2}}$ = 24%, $M_{\text{Gate 3}}$ = 44%) from the first gate on, and increased only somewhat between Gates 2 and 3.

In summary, the role of visual speech information for the recognition of spoken words varies over time. Most information provided by visual speech is fully available early on, whereas auditory information still accumulates. Visual speech, therefore, plays a more important role for recognition at the beginning of a phoneme than toward its end.

The Distribution of Visually Defined Featural Information

In the analyses so far, features were defined on the basis of linguistic theory that was originally developed to define distinctive linguistic features of auditorily perceived phonemes. Although these features also relate to differences in production, they may not be defined specifically enough to cover well the linguistic features transmitted by the face during word production. Therefore, %TI was also examined for a set of features that were explicitly based on more detailed articulatory distinctions (see Table 6). Note, however, that these features are still to be interpreted as linguistic features and not as lower level properties of the signals. That is, just as for the feature analyses reported previously, %TI scores reflect the overall transmission of these linguistic features and not what information in the signal contributes to their transmission.

Table 6
Feature Classification Scheme for Consonant Visemes and Their Included Phonemes

	Viseme/Phonemes											
Feature Classification	{p} {b,p,m}	{f} {f,v}	{th} {θ,ð}	{t} {t,d,n}	{k} {k,g}	{s} {s,z}	{ch} {ʃ,tʃ,dʒ}	{h} {h}	{j} {j}	{l} {l}	{r} {r}	{w} {w}
Duration	−	−	−	−	−	+	+	−	+	+	+	+
Tongue-tip movement	−	−	+	+	−	+	+	−	−	+	−	−
Lip rounding	−	−	−	−	−	−	−	−	+	−	+	+
Mouth narrowing	−	−	−	−	−	−	−	−	−	−	−	+
Dental adduction	−	−	+	−	−	+	−	−	+	−	−	−
Lower lip tuck	−	+	−	−	−	−	−	−	−	−	−	−
Protrusion	−	−	−	−	−	−	+	−	−	−	−	+
Labial closure	+	−	−	−	−	−	−	−	−	−	−	−

Note—"+" indicates the presence of a feature; "−" indicates the absence of a feature.

Consonants were classified by the features duration, tongue-tip movement, lip rounding, horizontal mouth narrowing, dental adduction, and lower lip tuck (C. S. Campbell & Massaro, 1997). Classifications for {ʃ}, {k}, and {h} were added. Lip rounding differed from its specification by Miller and Nicely (1955) in that, in the present study, {j} was also specified as rounded. For visemes with the linguistic feature "tongue-tip movement," the tip of the tongue visibly moves during their production. Mouth narrowing is a unique feature of the viseme {w}, for which the lips move horizontally closer during articulation. The dental adduction feature groups visemes by whether teeth are seen and are moving vertically closer during the uttering of the viseme. Lower lip tuck specifies a feature unique to the production of the viseme {f}, during which the lower lip is raised and placed underneath the upper front teeth. In addition, protrusion and labial closure were included to specify more visemes uniquely by their features. Only the visemes {k} and {h} share their feature specification. Protrusion distinguishes the palato-alveolar fricatives ({ch}) from laterals. Labial closure is a characteristic of the viseme that includes all bilabially produced plosives.

Figure 3 shows transmission values of all features in the three modality conditions over gates. The figure shows a generally higher transmission of features in the auditory than in the visual speech signal. Mouth narrowing (overall $M = 88\%$), lip rounding ($M = 87\%$), and labial closure ($M = 74\%$) information seems to be mostly available at the earliest gate. In the visual signal, it is noticeable that lower lip tuck appears to be mostly transmitted ($M = 39\%$). Information values for protrusion ($M = 15\%$), labial closure ($M = 12\%$), mouth narrowing ($M = 10\%$), and rounding ($M = 9\%$) were above 5% from the first gate on.

ANOVAs on the percentage of transmitted information for each feature were conducted with modality and gate as within-subjects factors. Table 7 gives an overview of these results. For all features, significant main effects of modality and gates were found (all $ps < .001$). The interaction between these two factors was significant for all features, with the exception of lower lip tuck.

Planned comparisons assessed the change in the percentage of transmitted information for each feature over the first three gates within each modality (see Table 8). In the auditory signal, the transmission of duration, tongue-tip movement, lip rounding, dental adduction, and protrusion increased over all gates. Values for mouth narrowing increased only between the first two gates, and for labial closure, only between the last two gates. Values for lower lip tuck did not change across gates. For the audiovisual presentation condition, the same relative pattern of featural information was generally found. The distribution of transmission values over gates also followed the auditory distribution in its pattern. One exception, however, is lower lip tuck, for which the amount of transmitted information significantly increased between Gates 2 and 3. In the visual signal, only information about tongue-tip movement and labial closure increased over the first two gates. There was a trend indicating an increase of lip rounding information over these gates as well. No other featural values improved between any two adjacent gates.

A second set of planned comparisons examined the featural distribution to the audiovisual benefit at each gate and its change over the first three gates (see Table 9). An audiovisual benefit at each gate was found for the features lower lip tuck (overall $M = 39\%$) and labial closure ($M = 30\%$). Lower lip tuck is a feature of labiodental fricatives. Labial closure is a feature of bilabial plosives. These features thus speak to the recognition of phonemes with a front place of articulation. The audiovisual benefit for lower lip tuck did not change across gates; the benefit for labial closure showed a numerical increase between Gates 1 and 2. Consonants with a mid place of articulation can be distinguished from other consonants by visual information on the linguistic feature tongue-tip movement. In the present study, visual speech contributed to a substantial audiovisual benefit at each gate ($M = 28\%$). This benefit somewhat improved over Gates 2 and 3. The distinction among visemes with a mid place of articulation was aided by visual speech information about the feature protrusion, which defines postalveolar fricatives and affricates (but which is also a characteristic of the labial-velar approximant {w}), as well as by information about lip rounding, which defines approximants (also by transmitting information about mouth narrowing for the labial-velar approximant {w}). The audiovisual benefit for protrusion ($M = 28\%$) and mouth narrowing ($M = 28\%$) was found at each gate. The benefit for protrusion somewhat improved over Gates 2 and 3, and the benefit for mouth narrowing between Gates 1 and 2. An audiovi-

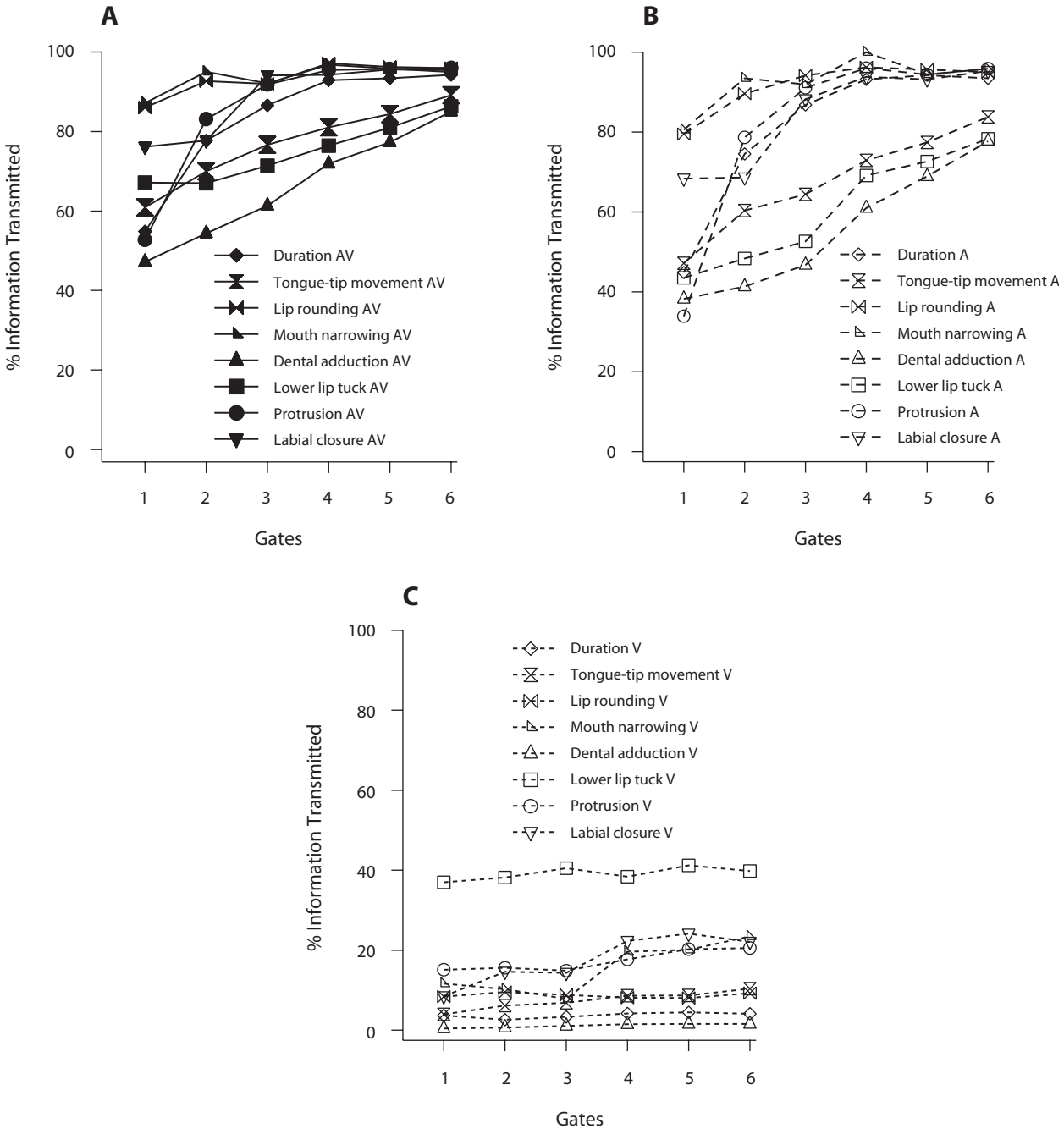

Figure 3. The percentage of information transmitted for the features duration, tongue-tip movement, lip rounding, mouth narrowing, dental adduction, lower lip tuck, protrusion, and labial closure based on consonant confusion data in (A) audiovisual, (B) auditory, and (C) visual speech over gates.

sual benefit for lip rounding (M = 23%) was found only at the first gate. At subsequent gates, this benefit was no longer statistically significant, but at the same time, it did not change substantially in size across gates. Information about dental and alveolar fricatives (and {j}) was given through the transmission of the feature dental adduction. Although relatively little information about this feature was contained in the visual signal, it led to a large benefit in audiovisual speech (M = 21%) at each gate, because of the relatively low feature values in the auditory signal. The audiovisual benefit for dental adduction increased only during Gates 2 and 3. An audiovisual benefit for duration that distinguishes fricative and affricative sibilants, but also approximants, from other phonemes was found only at the first gate (M = 16%). This benefit was no longer found at subsequent gates, even though a comparison of the size of the effect across gates did not reach significance.

Table 7
Results Overview of ANOVAs Conducted on the Percentage of Featural Information Transmitted for Visually Defined Features

	Modality			Gates			Modality × Gates		
Features	$F(2,220)$	p	η^2_G	$F(2,220)$	p	η^2_G	$F(4,440)$	p	η^2_G
Duration	4,248.74	<.001	.89	616.63	<.001	.46	232.50	<.001	.32
Tongue-tip movement	1,824.90	<.001	.81	179.25	<.001	.12	26.96	<.001	.04
Lip rounding	3,936.77	<.001	.90	44.83	<.001	.05	10.60	<.001	.03
Mouth narrowing	1,370.17	<.001	.80	8.75	<.001	.01	6.51	<.001	.02
Dental adduction	1,007.34	<.001	.76	81.94	<.001	.06	22.35	<.001	.03
Lower lip tuck	79.94	<.001	.21	17.63	<.001	.01	1.07*	.37	.002
Protrusion	1,576.10	<.001	.76	609.81	<.001	.46	238.44	<.001	.33
Labial closure	573.42	<.001	.65	90.46	<.001	.07	18.00	<.001	.02

Note—Degrees of freedom were Greenhouse–Geisser corrected (df_{effect} = 3.587; df_{error} = 394.57). *Sphericity assumption violated.

Table 8
Pairwise Comparisons of the Percentage of Information Transmitted of Visually Defined Consonantal Features Across the First Three Gates for Each Modality Condition

	Auditory Only				Audiovisual				Visual Only			
	Gates 1 vs. 2		Gates 2 vs. 3		Gates 1 vs. 2		Gates 2 vs. 3		Gates 1 vs. 2		Gates 2 vs. 3	
Features	t	p	t	p	t	p	t	p	t	p	t	p
Duration	20.94	<.001	9.47	<.001	15.81	<.001	7.53	<.001	−1.10	.27	0.27	.79
Tongue-tip movement	11.24	<.001	3.62	<.001	7.90	<.001	4.83	<.001	3.55	.001	0.33	.74
Lip rounding	6.77	<.001	2.55	.01	4.21	<.001	−0.41	.68	1.80	.08	−1.11	.27
Mouth narrowing	6.34	<.001	−0.56	.58	3.79	<.001	−1.64	.10	−0.49	.63	0.89	.37
Dental adduction	2.90	.005	4.11	<.001	6.04	<.001	5.59	<.001	1.36	.18	1.03	.30
Lower lip tuck	2.34	.02	2.21	.03	0.32	.75	3.31	.001	1.59	.12	0.71	.48
Protrusion	25.59	<.001	9.82	<.001	16.49	<.001	6.48	<.001	0.57	.57	−1.12	.27
Labial closure	−0.18	.86	10.09	<.001	0.70	.48	8.92	<.001	5.53	<.001	−0.72	.47

Note—The degree of freedom for all t tests was 110.

Table 9
Evaluations of the Relative Audiovisual Benefit of Transmitted Information of Visually Defined Consonantal Features for the First Three Gates, As Well As Pairwise Comparisons of the Size of This Benefit Across Gates

	Gates									
	1		2		3		1 vs. 2		2 vs. 3	
Features	t	p	t	p	t	p	t	p	t	p
Duration	6.04	<.001	−0.57	.57	1.53	.13	−0.65	.52	1.55	.12
Tongue-tip movement	10.61	<.001	7.51	<.001	10.57	<.001	0.24	.81	2.60	.01
Lip rounding	7.89	<.001	−0.54	.59	1.52	.13	−0.72	.47	1.54	.13
Mouth narrowing	11.45	<.001	10.78	<.001	10.89	<.001	−2.37	.02	1.32	.19
Dental adduction	7.76	<.001	5.42	<.001	9.88	<.001	1.55	.12	2.60	.01
Lower lip tuck	12.00	<.001	10.09	<.001	10.92	<.001	−0.56	.58	1.38	.17
Protrusion	10.66	<.001	5.94	<.001	8.20	<.001	0.14	.89	2.34	.02
Labial closure	5.13	<.001	9.63	<.001	13.07	<.001	1.70	.09	1.42	.16

Note—Positive t values indicate an audiovisual benefit and an increase of the audiovisual benefit across gates. The degree of freedom for all t tests was 110.

DISCUSSION

In the present large-scale study, we examined the temporal distribution of information in auditory and visual speech as it is used in unimodal and bimodal word recognition. The results demonstrated that visual and auditory speech differ not only in what featural information they provide, but also in when they do so during the speech signal. Visual speech information generally tends to be fully available early during the phoneme, whereas auditory speech information is accumulated across the phoneme. Hence, visual speech already has an impact on recognition early on. This difference in the temporal distribution of featural information in the modalities varies the degree to which features play a part in the audiovisual benefit while the speech signal unfolds.

Confusion patterns for gated auditory speech replicated the general patterns obtained from nongated syllables in noise (Benkí, 2003; Cutler, Weber, Smits, & Cooper, 2004; Miller & Nicely, 1955; Wang & Bilger, 1973). Auditory speech is generally the most informative about manner (mainly through rounding, nasality, continuant, and duration) and voicing, and the least informative about place of articulation. The basic pattern found for visual speech in

the present gating study replicates that commonly found in lip-reading studies of nongated American English consonants (see, e.g., Benguerel & Pichora-Fuller, 1982; Binnie, Montgomery, & Jackson, 1974; Owens & Blazek, 1985; Walden et al., 1974). Visual speech is mostly informative about place, is less so about manner, and contains little information about voicing.

These basic results are also in line with those found in a previous seminal Dutch audiovisual gating study that tracked information across gates of natural speech (Smeele, 1994), to the degree that that study addressed the same questions. The place of articulation in both studies was a highly visible feature from the first gate on, since it could be easily transmitted visually by its characteristic position and movements of lips, teeth, and the tongue (MacLeod & Summerfield, 1987; Summerfield, 1987). In the Dutch gating study, frication was the next highly visible feature, but this was shown to reflect mainly front place transmission, because of a confound in the Dutch materials. Our present study, without containing such a confound, showed a higher level of rounding than of frication information to be available visually early on. Rounding was, however, not analyzed in the Dutch gating study. In our present study, the additionally analyzed features continuant and duration were shown to be weakly transmitted visually. In both studies, there was virtually no visual nasality and voicing information available. This is not surprising, since voicing and nasality are produced by nonvisible articulators—namely, the vocal cord and the soft palate. We also examined the transmission of a set of visually defined distinctive linguistic features that showed that the auditory signal transmits mostly information about mouth narrowing, lip rounding, and labial closure early on. Visual speech also transmitted a fair amount of information about these three features as well as about protrusion. Lower lip tuck as a feature of labiodentals was, however, transmitted the most by visual speech.

Our present study and the Dutch gating study (Smeele, 1994), as well as the other gating studies (De la Vaux & Massaro, 2004; Munhall & Tohkura, 1998), showed that, generally, auditory and audiovisual phoneme and featural information increased similarly over gates, but that visual information increased less so, if at all. Audiovisual benefits for place (but never for voicing) were found early on and increased only slightly in size over gates. Despite methodological differences (e.g., fixed 40-msec gating of Dutch CV syllables) and the study's narrow scope, Smeele's (1994) study thus confirmed our basic and some more detailed results regarding the interplay of auditory and visual speech information while speech unfolds. Further detailed comparisons were not possible because of limitations in the scope of the Dutch study (e.g., limited materials) and its poorer temporal resolution.

The similarity between the more general results of our present study and those of the nongating literature, as well as Smeele's (1994) Dutch audiovisual gating study using a natural speaker, is important, since it confirms our virtual speaker as an appropriate choice. The speaker's visual speech and its temporal alignment to the auditory signal are based on natural speech motion data and their relation to the simultaneously produced auditory speech. These data were collected over the past two decades with a variety of different methods that are commonly used in articulatory phonetics (see, e.g., Cohen, Massaro, & Clark, 2002; for an overview, see Massaro, 1998). Production evaluations showed that the synthetic speech model replicates natural audiovisual speech dynamics (e.g., Cohen et al., 2002; Massaro, 1998). Converging evidence for the quality of the synthetic speech and its alignment comes also from a series of perceptual studies. These studies replicated audiovisual perceptual phenomena of a sensitive temporal nature, which was previously observed with natural audiovisual speech (see Massaro, 1998, for an overview). For example, the size of the McGurk effect and the nature of responses vary as a function of audiovisual temporal alignment (Munhall et al., 1996; Munhall & Tohkura, 1998). Our synthetic speaker produced McGurk effects that were similar in size and nature to those found with natural speakers. Likewise, studies with our synthetic speaker showed a similar sensitivity of the perceiver to temporal synchrony, as was found with natural speech (e.g., Grant & Greenberg, 2001; Massaro & Cohen, 1993; Massaro et al., 1996; Munhall et al., 1996). This would not have been the case if the two modalities were not synchronous in the synthetic speaker to begin with. This plentitude of empirical evidence from production and perception studies suggests the appropriateness of the speaker and his alignment for the purposes of the present study.

Nevertheless, we provide even more empirical support by once more directly comparing the synthetic speaker with a natural speaker within the same experiment. Ouni, Cohen, Ishak, and Massaro (2007) recently obtained confusion patterns of nine initial consonants under auditory, visual, and audiovisual presentations for the same version of a synthetic speaker that was used in the present study, as well as for a natural speaker who was known to have good intelligible visible speech (Bernstein & Eberhardt, 1986; Demorest & Bernstein, 1992; Lansing & McConkie, 2003). Auditory speech was presented under five noise levels, but to obtain a more robust data set, we pooled the confusion data across these noise conditions. To compare the synthetic and the natural speaker, we then calculated the correlation between all cells of the confusion matrices of both speakers. The correlation between all cells in the confusion matrix in audiovisual speech using the natural or the synthetic speaker was significant ($r = .93$; $p < .01$). A significant correlation was also found for visual confusion matrices ($r = .75$; $p < .01$). The speakers, therefore, provided highly similar information that was equally exploited by the perceivers to recognize speech. At the same time, similar information was lacking in both speakers, so that the errors made by perceivers were similar across speakers. This was the case for visual and audiovisual speech presentations. Hence, the alignment of synthetic visual speech to the auditory signal provides the same confusion patterns as those in the natural speaker. Unnatural temporal audiovisual asynchronies in the synthetic speech would have reduced these similarities. In conclusion, production and perceptual data—especially the similarity

between our gating results and those obtained by previous studies—suggest that the synthetic speaker was a suitable speaker for the present study. Our results seem, therefore, not to be limited in their scope by using audiovisual synthetic speech. Nonetheless, future studies should assess whether our results will replicate with a natural speaker. This replication may be necessary to further validate the visual speech of our talker and its alignment to the auditory signal at a finer resolution within a given phoneme. A further reason to replicate is provided by the well-known fact that talkers vary in their visual intelligibility (see, e.g., Jackson, 1988; Kricos & Lesner, 1982). Therefore, as is true for any one-talker study, future studies are also needed to examine whether our results hold across a series of natural speakers, or whether they vary as a function of talker intelligibility.

The main focus of the present study was to evaluate the temporal distribution of information across, but also within, modalities in audiovisual speech. Earlier research has shown that visual place information can precede auditory place information (De la Vaux & Massaro, 2004; Smeele, 1994). However, this has been demonstrated only in the limited case in which bilabial plosives were produced not only word initially, but also at the beginning of a speaker's turn. Normally, most sounds are produced in a continuous speech context. Then, coarticulatory information already contained in the preceding context cues the identity of an upcoming consonant (see, e.g., Smits et al., 2003). The present study showed that in this more natural case, there is generally early on more information contained in the auditory than in the visual signal. But the available visual information already occurs close to its maximal level of transmission early in the consonant, whereas the auditory information is more distributed across a phoneme. Visual speech information thus supplements early-on auditory information and leads to robust early recognition benefits, despite acoustic information that is already available. These early audiovisual benefits were not limited to the place of articulation, but were also found for rounding, frication, and duration, and for all visually defined linguistic features. Therefore, visual speech seems to play an important role during the early unfolding of a phoneme.

The availability of perceptual information over time differs between modalities and therefore contributes to a change in which features play a part in the audiovisual benefit while the speech signal unfolds. The audiovisual benefit varies not only within a word, but also within individual phonemes. An early benefit was found for the features place of articulation, rounding, frication, and duration. When the complete phoneme was presented, an audiovisual benefit was obtained for nasality and continuant, in addition to those found for place, rounding, and frication. However, although the benefits for nasality and continuant then reached significance, they were not substantially larger than the trends found for these two features on previous gates. For visually defined features, most contributed to an audiovisual benefit at all gates. Only duration and lip rounding were no longer contributing to an audiovisual benefit when all phoneme information became available. Only the audiovisual benefits for tongue-tip movement and labial closure increased in size over the first two gates. These two features reflect mid and front places of articulation. In summary, although an audiovisual phoneme recognition benefit can be found early on and does not vary in size over gates, visual speech information contributes differently to recognition at the beginning of a phoneme than it does toward the end. Visual and auditory speech not only differ in their featural informativeness, but they also do so differently over the speech signal and therefore vary the featural contribution to the audiovisual benefit over gates. Note, however, that auditory and audiovisual performance approach ceiling-level performance; thus, the absolute improvement because of visual speech decreases across a phoneme. The relative audiovisual benefit takes into account that for performance in the auditory condition approaching ceiling level, it becomes more difficult for visual speech to contribute an absolute improvement. The relative audiovisual benefit may not vary much across phonemes, but the absolute contribution of visual speech does. Hence, visual speech contributes to a larger absolute benefit at the beginning of a phoneme than it does toward its end.

Our present study adds to the previous literature by showing that visual and auditory speech differ not only in their informativeness across features, but also in what information they provide for a given feature. Visual speech provides mostly place information about the front rather than about the mid or back place. The visibility of front place information is reflected through the transmission of the features lower lip tuck for labiodental fricatives and labial closure for labial consonants. The mid and back places of articulation are difficult to distinguish, although the feature tongue-tip movements of consonants with a mid place can sometimes cue their places of articulation. Auditory speech contains more information about the back place rather than about the front or mid places of articulation. This leads to a higher transmission of front and back place than of mid place of articulation in the audiovisual signal. But visual and auditory speech not only provide different information for a given feature, they do so differently over time. Auditory and audiovisual mid and back place information increased over gates, but front place information did not. This result complements the distributions in the visual signal in which visual front and mid place information increased over the first two gates. Visual and auditory speech differ in what featural information becomes available when during the speech signal. This complementarity and redundancy of information across modalities over time led to audiovisual benefits for all three place features and place overall at all three gates, increasing only somewhat when the final parts of the phonemes were available.

The present study provides a starting point to investigate the effects of the temporal distribution of information on audiovisual word recognition. The general complementarity and redundancy of information in auditory and visual speech contribute to the efficiency and robustness of audiovisual word recognition. Because visual speech independently influences the lexical competitor space (Auer,

2002; Mattys et al., 2002) by providing redundant but also complementary featural information in audiovisual word recognition, it can rule out a different subset of potential competitors than can auditory speech. For example, visual place information should help one to recognize a target word *pin* and to rule out competitors *fin* or *kin*, but not the competitor *bin*. The place of articulation information is especially important for word recognition. Often, words differ only in the place of articulation of one of their constituent consonants (e.g., *pin* and *fin*; Greenberg, 2005). Also, it is the auditory place of articulation information that is the most vulnerable to noise and hearing impairment (Miller & Nicely, 1955; Summerfield, 1987). Therefore, visual speech information may play an important role in helping listeners recognize words that, on auditory grounds, would otherwise be difficult.

The results of the present study suggest that visual information can constrain word competition by providing information that rules out different competitors better than auditory information can, but that it also has the potential to do so earlier in time. For example, the audiovisual benefit is largely determined early on by the visual place of articulation information. Auditory information about this feature is still accumulated, whereas visual place information is already well transmitted early in the signal. The temporal variation in complementarity and redundancy provided by the audiovisual signals should influence the time course of audiovisual word recognition. Visual speech's contribution to phoneme recognition varies with the unfolding of a phoneme, but it does so differently across and within features. In absolute terms, seeing a speaker benefits recognition more at the beginning than at the end of a consonant. But also in terms of relative benefits, visual information plays a more important role during the early parts of the unfolding phoneme. The information is distributed across the two modalities differentially over time, and the nature of the benefit for audiovisual word recognition varies while the speech signal unfolds. The addition of visual speech thus not only generates predictions about the resolution of lexical ambiguity, but also about its temporal course. It is therefore critical to understand the time course by which information becomes available in the visual signal and its relationship to the availability of auditory information at any given point in time. The present study provides a good foundation for understanding the complex interplay of audiovisual information over time, and provides a valuable database for the testing and the development of quantitative models of multimodal word recognition.

AUTHOR NOTE

The present research was supported in part by doctoral grants from the University of California, Santa Cruz, to A.J. This work was part of the doctoral dissertation of A.J. Portions of this work have been presented in the proceedings of the Auditory–Visual Speech Processing International Conference, Parksville, Canada, July 2005. The authors thank James McQueen and three anonymous reviewers for discussion of earlier drafts. The authors thank Michael Cohen for help with creating stimuli recordings and with the model testing, as well as Shazia Bashiruddin, Astrid Carrillo, Martyna Citkowicz, Denise Coquia, Lisa Holton, and Neil Ryan for the testing of the participants. A.J. is now at the Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands. Address correspondence to A. Jesse, Max Planck Institute for Psycholinguistics, Postbus 310, 6500AH Nijmegen, The Netherlands (e-mail: alexandra.jesse@mpi.nl).

REFERENCES

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory & Language*, **38**, 419-439. doi:10.1006/jmla.1997.2558

Auer, E. T., Jr. (2002). The influence of the lexicon on speech read word recognition: Contrasting segmental and lexical distinctiveness. *Psychonomic Bulletin & Review*, **9**, 341-347.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, **37**, 379-384.

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

Benkí, J. R. (2003). Analysis of English nonsense syllable recognition in noise. *Phonetica*, **60**, 129-157. doi:10.1159/000071450

Bernstein, L. E., & Eberhardt, S. P. (1986). *Johns Hopkins lipreading corpus I–II: Disc I* [Videodisk]. Baltimore: Johns Hopkins University.

Binnie, C. A., Montgomery, A. A., & Jackson, P. L. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech & Hearing Disorders*, **17**, 619-630.

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **30**, 445-463. doi:10.1037/0096-1523.30.3.445

Breeuwer, M., & Plomp, R. (1986). Speechreading supplemented with auditorily presented speech parameters. *Journal of the Acoustical Society of America*, **79**, 481-499. doi:10.1121/1.393536

Campbell, C. S., & Massaro, D. W. (1997). Perception of visible speech: Influence of spatial quantization. *Perception*, **26**, 627-644. doi:10.1068/p260627

Campbell, R., & Dodd, B. (1980). Hearing by eye. *Quarterly Journal of Experimental Psychology*, **32**, 85-99. doi:10.1080/00335558008248235

Cathiard, M.-A., Lallouache, M. T., Mohamadi, T., & Abry, C. (1995). Configurational vs. temporal coherence in audio-visual speech perception. In K. Elenius & P. Branderud (Eds.), *Proceedings of the 13th International Congress of Phonetic Sciences* (Vol. 3, pp. 218-221). Stockholm: ICPhS.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York: Harper & Row.

Cohen, M. M., Massaro, D. W., & Clark, R. (2002). Training a talking head. In D. C. Martin (Ed.), *Proceedings of the IEEE Fourth International Conference on Multimodal Interfaces* (ICMI'02) (pp. 499-510). Pittsburgh. doi:10.1109/ICMI.2002.1167046

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *Journal of the Acoustical Society of America*, **116**, 3668-3678. doi:10.1121/1.1810292

Davis, M. H., Marslen-Wilson, W. D., & Gaskell, M. G. (2002). Leading up the lexical garden path: Segmentation and ambiguity in spoken word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **28**, 218-244.

De la Vaux, S. K., & Massaro, D. W. (2004). Audiovisual speech gating: Examining information and information processing. *Cognitive Processing*, **5**, 106-112. doi:10.1007/s10339-004-0014-2

Demorest, M. E., & Bernstein, L. E. (1992). Sources of variability in speechreading sentences: A generalizability analysis. *Journal of Speech & Hearing Research*, **35**, 876-891.

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

Fisher, C. G. (1968). Confusions among visually perceived consonants. *Journal of Speech & Hearing Research*, **15**, 474-482.

Grant, K. W., & Greenberg, S. (2001). Speech intelligibility derived from asynchronous processing of auditory-visual information. In D. W. Massaro, J. Light, & K. Geraci (Eds.), *Proceedings of Auditory–Visual Speech Processing (AVSP-2001)* (pp. 132-137). Santa Cruz, CA: Perceptual Science Laboratory.

Grant, K. W., & Walden, B. E. (1996). Evaluating the articulation index

for audio-visual consonant recognition. *Journal of the Acoustical Society of America*, **100**, 2415-2424. doi:10.1121/1.417950

GREENBERG, S. (1999). Speaking in shorthand—A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, **29**, 159-176. doi:10.1016/S0167-6393(99)00050-3

GREENBERG, S. (2005). A multi-tier framework for understanding spoken language. In S. Greenberg & W. Ainsworth (Eds.), *Listening to speech: An auditory perspective* (pp. 411-433). Hillsdale, NJ: Erlbaum.

GRIMM, W. A. (1966). Perception of segments of English-spoken consonant-vowel syllables. *Journal of the Acoustical Society of America*, **40**, 1454-1461. doi:10.1121/1.1910248

GROSJEAN, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, **28**, 267-283.

GROSJEAN, F. (1996). Gating. *Language & Cognitive Processes*, **11**, 597-604. doi:10.1080/016909696386999

JACKSON, P. L. (1988). The theoretical minimal unit for visual speech perception. *Volta Review*, **90**, 99-115.

KIEFTE, M. (2003). Temporal information in gated stop consonants. *Speech Communication*, **40**, 315-333. doi:10.1016/S0167-6393(02)00069-9

KRICOS, P. B., & LESNER, S. A. (1982). Differences in visual intelligibility across talkers. *Volta Review*, **84**, 219-225.

KUČERA, H., & FRANCIS, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

LANSING, C. R., MCCONKIE, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics*, **65**, 536-552.

LISKER, L., & ROSSI, M. (1992). Auditory and visual cueing of the [± rounded] feature of vowels. *Language & Speech*, **35**, 391-417.

LÖFQVIST, A. (1990). Speech as audible gestures. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modeling* (pp. 289-322). Dordrecht, NL: Kluwer.

MACLEOD, A., & SUMMERFIELD, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, **21**, 131-141. doi:10.3109/03005368709077786

MAGNUSON, J. S., DIXON, J., TANENHAUS, M. K., & ASLIN, R. N. (2007). The dynamics of lexical competition during spoken word recognition. *Cognitive Science*, **31**, 133-156. doi:10.1207/s15516709cog3101_5

MASSARO, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle.* Cambridge, MA: MIT Press.

MASSARO, D. W., & COHEN, M. M. (1993). Perceiving asynchronous bimodal speech in consonant–vowel and vowel syllables. *Speech Communication*, **13**, 127-134. doi:10.1016/0167-6393(93)90064-R

MASSARO, D. W., COHEN, M. M., & SMEELE, P. M. T. (1996). Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America*, **100**, 1777-1786. doi:10.1121/1.417342

MATTYS, S. L., BERNSTEIN, L. E., & AUER, E. T., JR. (2002). Stimulus-based lexical distinctiveness as a general word-recognition mechanism. *Perception & Psychophysics*, **64**, 667-679.

MCGURK, H., & MACDONALD, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748. doi:10.1038/264746a0

MILLER, G. A., & NICELY, P. A. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, **27**, 338-352. doi:10.1121/1.1907526

MUNHALL, K. G., GRIBBLE, P., SACCO, L., & WARD, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics*, **58**, 351-362.

MUNHALL, K. G., & TOHKURA, Y. (1998). Audiovisual gating and the time course of speech perception. *Journal of the Acoustical Society of America*, **104**, 530-539. doi:10.1121/1.423300

NORRIS, D., & MCQUEEN, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, **115**, 357-395. doi:10.1037/0033-295X.115.2.357

ÖHMAN, S. E. G. (1966). Perception of segments of VCCV utterances. *Journal of the Acoustical Society of America*, **40**, 979-988. doi:10.1121/1.1910222

OLEJNIK, S., & ALGINA, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, **8**, 434-447. doi:10.1037/1082-989X.8.4.434

OUNI, S., COHEN, M. M., ISHAK, H., & MASSARO, D. W. (2007). Visual contribution to speech perception: Measuring the intelligibility of animated talking heads. *EURASIP Journal on Audio, Speech, & Music Processing*, **2007** (Art. No. 47891). doi:10.1155/2007/47891

OWENS, E., & BLAZEK, B. (1985). Visemes observed by hearing-impaired and normal-hearing adult viewers. *Journal of Speech & Hearing Research*, **28**, 381-393.

POLS, L. C. W., & SCHOUTEN, M. E. H. (1978). Identification of deleted consonants. *Journal of the Acoustical Society of America*, **64**, 1333-1337. doi:10.1121/1.382100

REISBERG, D., MCLEAN, J., & GOLDFIELD, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97-113). Hillsdale, NJ: Erlbaum.

ROBERT-RIBES, J., SCHWARTZ, J.-L., LALLOUACHE, T., & ESCUDIER, P. (1998). Complementarity and synergy in bimodal speech: Auditory, visual and audio-visual identification of French oral vowels in noise. *Journal of the Acoustical Society of America*, **103**, 3677-3689. doi:10.1121/1.423069

SALTZMAN, E. L., & MUNHALL, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, **1**, 1615-1623. doi:10.1207/s15326969eco0104_2

SEITZ, P. F., & GRANT, K. W. (1999). Modality, perceptual encoding speed, and time-course of phonetic information. In D. W. Massaro (Ed.), *Proceedings of Auditory–Visual Speech Processing (AVSP'99)*. Santa Cruz, CA.

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

SMEELE, P. M. T. (1994). *Perceiving speech: Integrating auditory and visual speech*. Unpublished doctoral dissertation, Delft University of Technology, The Netherlands.

SMITS, R. (2000). Temporal distribution of information for human consonant recognition in VCV utterances. *Journal of Phonetics*, **27**, 111-135. doi:10.1006/jpho.2000.0107

SMITS, R., WARNER, N., MCQUEEN, J. M., & CUTLER, A. (2003). Unfolding of phonetic information over time: A database of Dutch diphone perception. *Journal of the Acoustical Society of America*, **113**, 563-574. doi:10.1121/1.1525287

STEVENS, K. N. (2000). *Acoustic phonetics*. Cambridge, MA: MIT Press.

STEVENS, K. N., & BLUMSTEIN, S. E. (1978). Invariant cues for place of articulation in stop consonants. *Journal of the Acoustical Society of America*, **64**, 1358-1368. doi:10.1121/1.382102

SUMBY, W. H., & POLLACK, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, **26**, 212-215. doi:10.1121/1.1907309

SUMMERFIELD, A. Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3-52). Hillsdale, NJ: Erlbaum.

TEKIELI, M. E., & CULLINAN, W. L. (1979). The perception of temporally segmented vowels and consonant–vowel syllables. *Journal of Speech & Hearing Research*, **22**, 103-121.

TRAUNMÜLLER, H., & ÖHRSTRÖM, N. (2007). Audiovisual perception of openness and lip rounding in front vowels. *Journal of Phonetics*, **35**, 244-258. doi:10.1016/j.wocn.2006.03.002

TYLER, L. K. (1984). The structure of the initial cohort: Evidence from gating. *Perception & Psychophysics*, **36**, 417-427.

VAN WASSENHOVE, V. (2004). *Cortical dynamics of auditory-visual speech: A forward model of multisensory integration*. Unpublished doctoral dissertation, University of Maryland.

WALDEN, B. E., PROSEK, R. A., & WORTHINGTON, D. W. (1974). Predicting audiovisual consonant recognition performance of hearing-impaired adults. *Journal of Speech & Hearing Research*, **17**, 270-278.

WANG, M. D., & BILGER, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *Journal of the Acoustical Society of America*, **54**, 1248-1266. doi:10.1121/1.1914417

YEHIA, H., RUBIN, P., & VATIKIOTIS-BATESON, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication*, **16**, 23-43. doi:10.1016/S0167-6393(98)00048-X

ZWITSERLOOD, P. (1989). The locus of the effects of sentential–semantic context in spoken-word processing. *Cognition*, **32**, 25-64. doi:10.1016/0010-0277(89)90013-9

APPENDIX
Experimental Stimuli With Phoneme Duration Information (Initial Consonant Duration and Vowel Duration, in Milliseconds)

bath (78, 268), beef (75, 192), book (71, 135), pack (91, 250), peak (96, 145), push (131, 130), match (67, 180), moth (84, 209), mood (125, 196), fetch (173, 148), foot (191, 129), fish (154, 178), vat (64, 183), veal (60, 129), veer (60, 98), that (36, 177), these (61, 241), them (48, 139), theme (149, 167), though (126, 214), thick (105, 171), teach (114, 224), tall (123, 246), tomb (114, 162), dog (76, 309), deck (76, 218), dig (52, 169), nap (68, 157), nod (47, 190), noon (83, 132), cause (100, 265), cash (109, 248), cook (106, 121), gang (113, 138), good (100, 183), give (115, 151), sad (151, 244), sock (110, 182), sing (133, 63), czar (134, 165), zeal (113, 161), zip (116, 105), sheep (190, 153), shed (123, 167), shoot (159, 144), chop (136, 183), choose (130, 203), chair (170, 118), job (68, 249), jazz (141, 287), juice (107, 161), hawk (100, 203), hen (147, 83), hood (97, 142), yawn (37, 153), yell (85, 60), youth (106, 127), leg (162, 187), loop (119, 102), lid (97, 118), roar (61, 83), roof (132, 133), rich (53, 134), wash (115, 201), wet (105, 178), weave (82, 186).

(Manuscript received January 9, 2008;
revision accepted for publication July 28, 2009.)