Visual Phonemic Ambiguity and Speechreading

Björn Lidestam
Linköping University, Linköping, Sweden

Jonas Beskow
Centre for Speech Technology, KTH, Stockholm, Sweden

Purpose: To study the role of visual perception of phonemes in visual perception of sentences and words among normal-hearing individuals.
Method: Twenty-four normal-hearing adults identified consonants, words, and sentences, spoken by either a human or a synthetic talker. The synthetic talker was programmed with identical parameters within phoneme groups, hypothetically resulting in simplified articulation. Proportions of correctly identified phonemes per participant, condition, and task, as well as sensitivity to single consonants and clusters of consonants, were measured. Groups of mutually exclusive consonants were used for sensitivity analyses and hierarchical cluster analyses.
Results: Consonant identification performance did not differ as a function of talker, nor did average sensitivity to single consonants. The bilabial and labiodental clusters were most readily identified and cohesive for both talkers. Word and sentence identification was better for the human talker than the synthetic talker. The participants were more sensitive to the clusters of the least visible consonants with the human talker than with the synthetic talker.
Conclusions: It is suggested that ability to distiguish between clusters of the least visually distinct phonemes is important in speechreading. Specifically, it reduces the number of candidates, and thereby facilitates lexical identification.

KEY WORDS: speechreading, articulation, students, normal hearing

For normal-hearing persons under normal listening conditions, speech perception and understanding are effortless and accurate. If the acoustic or sensory information is degraded (e.g., by noise or hearing impairment) seeing the talker's speech movements can compensate for the loss of auditory information, because the auditory and visual speech signals complement each other well: features of speech that are difficult to hear in noise are relatively easy to identify visually, and vice versa (Summerfield, 1983).

However, it is much harder to perceive and understand what someone is saying without hearing when you can just see the speech movements and not hear what they say, as the case is in speechreading. The difficulty lies in the fact that the visual speech signal is poorly specified for a number of reasons. First, some phonemes have features that normally are hidden from sight. For example, the vibrations of vocal cords, which distinguish voiced consonants from unvoiced consonants, are not visible (Lisker & Abramson, 1964), since we usually do not see the vocal cords. Second, those phonemes that can be seen relatively easily are often very difficult to distinguish from each other, as the places of articulation may be very closely located. Some phonemes that share visual articulatory characteristics are easily confused when they are presented in a visual-only modality, and these groups of easily confused phonemes are sometimes

referred to as visemes (Berger, 1972; Summerfield, 1983; van Son, Huiskamp, Bosman, & Smoorenburg, 1993; Walden, Prosek, Montgomery, Scherr, & Jones, 1977).

In spite of difficulties associated with the visual identification of phonetic information, some individuals can speechread with astonishing accuracy. Reported cases of extremely proficient speechreaders all concern persons who are hearing impaired or deaf (Andersson & Lidestam, 2005; Lyxell, 1994; Rönnberg, 1993). Group data have also revealed better speechreading for severely hearing-impaired and deaf individuals than for participants with normal hearing (Bernstein, Demorest, & Tucker, 2000; Ellis, MacSweeney, Dodd, & Campbell, 2001). On the basis of reported case studies of extreme speechreading skill up to that point in time, Rönnberg, Samuelsson, and Lyxell (1998) proposed that extreme speechreading capacity can only be obtained if the speechreader has superior working memory capacity and uses some higher level, top-down processing strategies (i.e., that speechreading is driven by expectations). However, Andersson and Lidestam (2005) reported a case study of an expert speechreader who neither proved to have superior working memory capacity nor reported relying excessively on top-down processing. Instead, superior bottom-up capacity in the form of excellent phoneme identification, coupled with excellent executive functions, formed the basis for bottom-up driven speechreading. Thus, sensitivity to phonemes is a key factor in speechreading. This sensitivity may interact with lexical constraints in word and sentence identification (Auer, 2002; Auer & Bernstein, 1996; Auer, Bernstein, & Mattys, 2001; Mattys, Bernstein, & Auer, 2002).

The general purpose of this study was to investigate the role of phoneme identification in the visual perception of sentences and words among individuals with normal hearing. Perception of single phonemes, spoken without a linguistic context, is not influenced to any great extent by top-down processing strategies, since the phonemes by themselves are devoid of semantic information. Perception of words and sentences, on the other hand, may be highly dependent on top-down processing strategies. The complementary information, which can be used for top-down processing strategies, may come from various sources, including linguistic, topical, and paralinguistic context (e.g., Lidestam, Lyxell, & Andersson, 1999; Marslen-Wilson, 1995; Samuelsson & Rönnberg, 1993). In this report, complementary information is defined as all information that is provided prior to or at the same time as the phonetic signal features, and that may constitute cues to what will be uttered or is being uttered. For example, this may mean knowing that the person you see talking is your doctor and seeing her smile when she says "you will be well in no time." Such topical and emotional cues may help to disambiguate semantically ambiguous words (cf. Rodd, Gaskell, & Marslen-Wilson, 2004) or to disambiguate a poorly specified speech signal (cf. Rönnberg et al., 1998).

Bernstein and colleagues (e.g., Bernstein, Demorest, & Eberhardt, 1994; Bernstein et al., 2000; Bernstein, Iverson, & Auer, 1997; Demorest, Bernstein, & DeHaven, 1996; Iverson, Bernstein, & Auer, 1998) have focused on bottom-up processing and stressed that the ability to extract as much information as possible from the visual speech signal is crucial to speechreading. Rönnberg and colleagues (e.g., Rönnberg, 1995; Rönnberg, Arlinger, Lyxell, & Kinnefors, 1989; Rönnberg et al., 1998; Samuelsson & Rönnberg, 1993) have focused on top-down processing of complementary information and stated that higher order cognitive functions, such as working-memory capacity, are important for speechreading proficiency. The present study incorporates measures of accuracy in the perception of, and sensitivity to, the signal and at the same time takes into consideration that the perception of the signal may or may not be influenced by additional information such as topical and emotional cues.

The first specific purpose of the present study was to assess and describe how accurately individuals with normal hearing perceive phonemic information visually under naturalistic conditions. This means that phonemes were identified without training, and the participants were not given feedback after each response. In addition, emotional facial expressions were included—as is the case in everyday life. Phoneme identification with variation of the talker's emotional facial expressions has not been studied previously and there is no evidence of correlation between identification of phonemes on the one hand and topically or emotionally cued identification of words and sentences on the other hand. In the present study, analysis was performed with signal detection methodology (Green & Swets, 1966; Macmillan & Creelman, 1991). Sensitivity to single consonants as well as to clusters of consonants was assessed, and hierarchical cluster analysis (Sneath & Sokal, 1973) was used to explore patterns of confusions among the consonants (cf. visemes).

The second specific purpose of the study was to investigate how consonant identification, which is devoid of semantic information, is correlated with word and sentence identification, which entail ample opportunity for complementing information to affect perception via top-down processing strategies. How much of the variation that is accounted for by bottom-up processing in different analytical levels of speechreading (i.e., sentences vs. single, short words) with different levels of additional information (i.e., with vs. without topical information; with vs. without emotional information) was therefore investigated. Bernstein et al. (2000) reported significant but low correlations ($r = .40$ to $r = .43$) between phoneme (/Ca/)

Table 1. Mean percentages and ranges for the consonant identification task by talker and displayed emotion.

	Talker			
	Human		Synthetic	
Displayed emotion	*M*	Range	*M*	Range
Neutral	22	8–36	23	8–53
Positive	24	14–39	23	11–64
Negative	16	3–25	23	3–61

Note. Performance was scored as proportion of correctly identified consonants.

identification scores and sentence and word identification scores obtained from normal-hearing participants, using the same (male) talker and the same dependent measure (proportion phonemes correct). However, the relationship between phoneme identification and semantically cued speechreading with topical and emotional cues has not been investigated.

The final specific purpose was to explore what aspects of visual phonemic information are most important for visual perception of words and sentences. To assess the effects of ambiguous visual phonemic information (i.e., when the quality or distinctiveness of articulation is poor), a synthetic talker (Beskow, 1997) was used along with a human talker. The synthetic talker has identical default parameters for many phonemes, including the following sets of consonants under scrutiny in this study: /b m/, /d n t/, /f v/, /h k ŋ ɧ/, and /r l/ (Beskow, 1997). Identical default parameters may result in the loss of subtle phonemic features available in natural (human) visual speech. For example, a human may inflate his or her cheeks when pronouncing a /b/, but not when pronouncing an /m/. The present synthetic talker, however, does not have a parameter for cheek inflation and has identical default parameters for /b m/. Lidestam, Beskow, and Lyxell (2006) found that it seems that the synthetic talker articulates consonants as distinctly as a human talker, judging by the proportion of correctly identified consonants. Expressed in the t statistic, the difference was $t(23) = 1.17$, *ns,* $d = .30$. Table 1 further specifies performance level as a function of displayed emotion.

However, mean scores are lower for the synthetic talker than for the human talker in word identification, $t(23) = 11.78$, $p < .001$, $d = 2.55$, and in sentence identification, $t(23) = 6.93$, $p < .001$, $d = 1.71$ (Lidestam et al., 2006). Table 2 further specifies performance levels as functions of topic and emotion. This indicates that the difference between the human and the synthetic talker in either articulation or coarticulation, or both, affects phoneme perception. This effect is apparent in the word and sentence identification tasks, but not in the consonant identification task. Therefore, further analyses of differences in accuracy in the perception of, and sensitivity to, phonemic information with the natural (human) talker and the synthetic talker were made in the present study. Comparison of phoneme perception using talkers who seem to generate different levels of phonemic ambiguity allowed further insights into the aspects of the phonemic information that are most important for the perception of the linguistically more complex words and sentences. Therefore, effects of talker were also studied in conjunction with the first and second purposes.

Table 2. Mean percentages and ranges for the word and sentence identification tasks by talker, topic, and emotion.

		Emotion					
		No cue		Displayed		Cue word	
Talker	Linguistic complexity	*M*	Range	*M*	Range	*M*	Range
Without topical cue							
Human	Word	35	11–67	33	13–61	38	7–75
	Sentence	26	8–50	24	8–49	27	9–55
Synthetic	Word	19	3–30	20	3–39	20	6–57
	Sentence	14	5–30	14	5–29	16	7–34
With topical cue							
Human	Word	33	3–58	38	3–58	38	11–100
	Sentence	29	7–60	32	7–78	33	8–100
Synthetic	Word	17	3–32	21	3–54	19	3–45
	Sentence	19	7–45	18	7–33	20	5–53

Note. Performance was scored as proportion of correctly identified consonants in correct serial position per item.

Method

Participants

Twenty-four students with normal hearing, 7 of them male, were paid 50 SEK (approximately $7) for participation. Ages ranged between 19 and 40 years (M = 24.4, SD = 3.2), and participants reported having Swedish as their native language, normal or corrected visual acuity, no hearing loss, and no prior training in speechreading.

Design

For the consonant identification task, the design was 3 × 2 factorial, the first variable being displayed emotion (neutral, positive, and negative) and the second being talker (human talker and synthetic talking head). Both variables were within groups.

For the sentence and word identification tasks, a 2 × 2 × 2 × 3 factorial design was used. The first variable was linguistic level (sentence vs. word identification), the second was talker (human talker and synthetic talking head), the third was topic (topical cue words or not), and the fourth was emotion (no emotional cues, displayed emotion, and emotional cue words). All variables were within groups. The sentence and word identification tasks each consisted of 12 blocks. Each block consisted of six items and represented a cell of the conditions, such as "no topical cue, human talker, and displayed emotion." The presentation order of topic, talker, and emotion was balanced, resulting in 12 (2 × 2 × 3) presentation orders, each comprising 2 participants. Talker constituted the largest coherent block of the balanced within-groups variables, topic the second largest, and emotion the smallest. The order of talker, topic, and emotion was the same within individuals throughout all tasks. The order of items was fixed (see the Appendix for examples), in order to assign each item to all conditions. The order of tasks was also fixed: first, the sentence identification task; second, the consonant identification task; and finally, the word identification task.

Materials

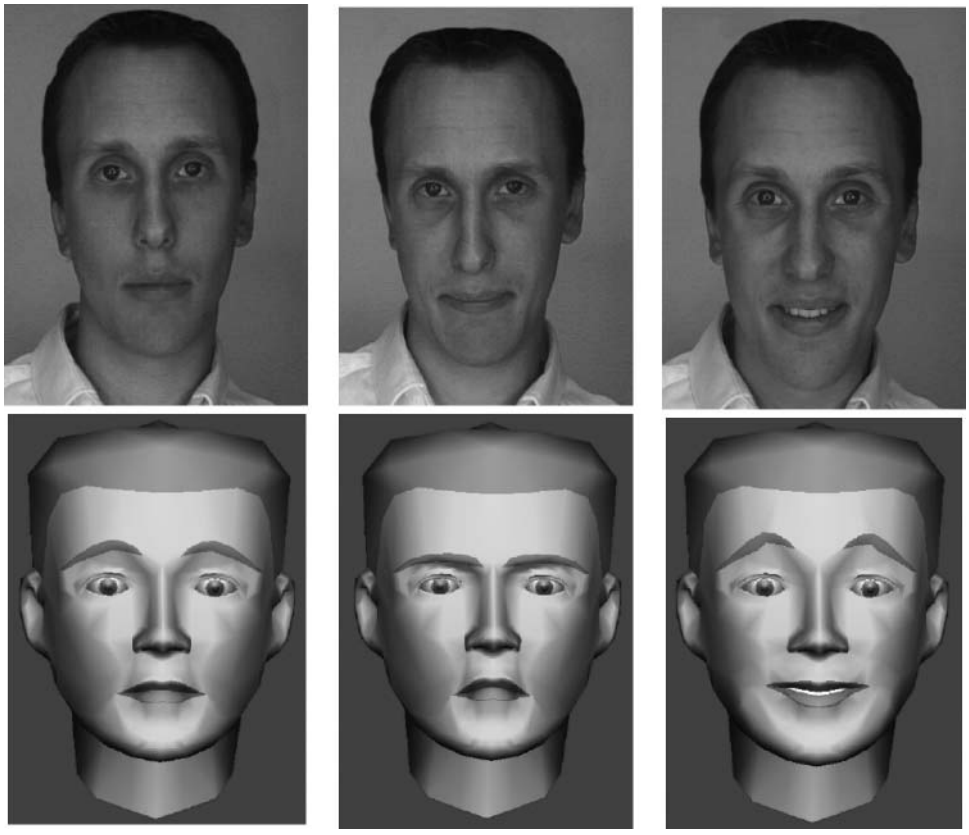
Apparatus and settings. A Sony DV Cam DSR-200P was used for the video recordings. A PC with an Intel Pentium 4 processor (1.8 GHz clock frequency), a Pinnacle Studio DV card, an NVIDIA GeForce2 MX/MX 400 video card, and a 17 in. high-resolution monitor set at 1024 × 768 pixels and an 85 Hz refresh rate were used for editing and presenting the stimuli. The frame rate of the video recordings of the human talker was 25 frames per second, with a resolution of 720 × 576 pixels. A Brüel & Kjær Type 2205 sound level meter and a Brüel & Kjær Type 4117 1 in. free-field microphone were used to monitor noise levels. Broadband noise was used for the purpose of allowing comparisons with the auditory and audiovisual conditions in Lidestam et al. (2006), where noise was used to make the auditory speech signal ambiguous.

The synthetic talker was a parametrically controlled three-dimensional polygonal model (Beskow, 1997, see Figure 1) that was animated in synchrony with natural speech from the video recordings. The model has parameters for speech articulation as well as for facial expressions. There are seven parameters controlling the articulation: jaw rotation, labiodental occlusion, bilabial occlusion, lip rounding, lip protrusion, mouth spread, and tongue-tip elevation. The articulatory parameters are controlled by a set of rules that map the phonetic transcription and associated durations of the acoustic speech into continuous trajectories, taking coarticulation into account (Beskow, 1995). The speech material was phonetically transcribed using the transcription module of the KTH Text-to-Speech system (Carlson, Granström, & Hunnicutt, 1982). The phonetic transcriptions were then automatically aligned with the audio signal from the .wav files using the HTK Toolkit (Young, Odell, Ollason, Valtchev, & Woodland, 1997) with talker adaptation of the acoustic models. The alignments were thereafter checked. The resulting time-aligned phoneme sequences were used as input to the visual speech synthesis rule system to generate the articulatory parameter trajectories that drove the synthetic talker. Further, eight emotional expressions (plus neutral facial expression) were used as analogies to the emotional cue words (i.e., happy, sad, disappointed, stern, concerned, disgusted, angry, and afraid; see Lidestam et al., 2006). The emotional expression parameters were brow raising, brow frown, eye opening, smile, and gaze. The emotional and articulatory parameters are independent; emotional expressions do not affect the articulation, or vice versa. Lidestam et al. verified that both talkers distinctly conveyed all three levels of emotional valence (i.e., neutral, positive, and negative) in the consonant identification task. The synthetic talker also distinctly displayed all three levels of valence in the word and sentence identification tasks, and the human talker conveyed distinct positive valence, but no distinct difference between neutral and negative valence (Lidestam et al.).

The participants were seated in front of the screen, about 60 cm from the monitor, and 30 cm from each other. A screen prevented them from seeing each other's reply sheets. The noise level was 62.5 dB (A) at the points of the participants' ears.

Preparation of stimulus materials. All stimuli were video recordings of a male actor who was hired due to his reputation of having vivid emotional expressions. Lighting prevented shadows on the actor's face, and he was clean-shaven in order to optimize speechreading. The

Figure 1. The human and the synthetic talker displaying neutral, negative (angry), and positive (happy) emotion.



area from the top of the actor's head to the top of his shoulders was recorded full-face (see Figure 1). The actor spoke the sentences as naturally as possible, but with constant rhythm, speed, and volume throughout (i.e., the variation was in displayed emotion). Thus, the effect of displayed emotion was intended to be visually distinct, and auditorily distinguishable only as variation in pitch and intonation. The DV recording was saved in separate noncompressed .avi files. The soundtracks from these .avi files were exported as .wav files and used for the synthetic talker condition.

Consonant identification task. The 18 most frequently occuring consonants in the word and sentence identification tasks /b d f g h j k l m n p r s t v ŋ ɧ ʈ/ were chosen for the consonant identification task. These consonants were presented between pairs of the vowel /a/ (i.e., /aCa/). There were 216 items: 18 (consonants) × 3 (displayed emotions) × 2 (talkers) × 2 (replications). That is, each unique recording of an /aCa/ stimulus was presented twice using the same talker with the same displayed emotion (e.g., the human talker displaying happiness while speaking /aba/). All items were preceded by a number indicating its serial order, for 0.5 s. The items were arranged in order for the participants to get the impression that the presentation order was randomized. Performance was scored as proportion of correct responses.

Word and sentence identification tasks. Seventy-two sentences of varying length (2–9 words; 7–33 phonemes) and 72 mono- and bisyllabic words (2–6 phonemes) from Lidestam et al. (1999) constituted the word and sentence identification tasks. The word and sentence identification tasks started with a "visit-to-a-restaurant" scenario, followed by a "visit-to-the-doctor" scenario. All sentences had been rated high on valence and typicality for these scenarios in Lidestam et al. The blocks of sentences and words consisted of six items each; three with positive and three with negative valence. These blocks were matched with regard to sentence length, word length, and difficulty of items (Lidestam, 2002). In the emotional cue word condition, only stimuli with neutral displayed emotion were used.

Emotion conveyed via emotional cue words was restricted to the following eight words: *happy* (for positive valence), *sad, disappointed, stern, concerned, disgusted, angry,* and *afraid* (for negative valence). These emotional cue words were chosen so that the emotional meaning of the message would be conveyed as clearly as possible. The same cues were used by the actor for generating

emotional facial expressions. Seven negative cue words and one positive cue word were chosen to reflect the fact that the Swedish language has more one-word expressions for negative valence than for positive valence.

Mean performance levels. The phrases *visit to a doctor* and *visit to a restaurant* were use to cue topic. Information about these scenarios was given both in writing and orally by way of introduction. Performance was scored as proportion of correctly identified phonemes in correct serial position (Lidestam, 2002).

Procedure

The trials were run with either 1 or 2 participants at a time. The participants wrote their responses on reply sheets. The participants were encouraged to guess and were told not to leave blanks. The experiment was run without practice trials. Presentation rate was adjusted to the participants' response rates.

Cue words were presented (e.g., *visit to a restaurant – disappointed*) as black text on a white background for 3 s before the first frame of the stimulus appeared. The first frame of each stimulus was frozen for 2 s before the stimulus was played, and the last frame appeared frozen for 3 s. Then the screen turned black until the experimenter pressed a key and a new cue word (or a stimulus without prior cue words) was presented.

Before each new type of task (i.e., the sentence, consonant, and word identification task, in serial order) and cueing (i.e., topical and emotional cue words), instructions were given both orally and as text on a sheet of paper. The reply sheets had numbered lines for the word and sentence identification tasks and numbered multiple-choice matrices with all consonants for the consonant identification task. The test sessions lasted 71–118 min. There was a short break for coffee, tea, bread, and biscuits after the sentence identification and consonant identification tasks.

Results and Discussion

Consonant identification. Table 1 presents the means and ranges for the consonant identification task by talker and displayed emotion. A repeated measures factorial analysis of variance (ANOVA) for the consonant identification task yielded an interaction between talker and displayed emotion, $F(2, 46) = 5.76$, $p < .01$. There was a simple effect of displayed emotion for the human talker, $F(2, 92) = 10.70$, $p < .001$. Multiple comparisons revealed that the effect was due to lower performance for the negative displayed emotion condition ($p < .01$). This indicates that consonants were more difficult to identify in the human talker when he displayed negative emotion, possibly because he used less distinct articulation patterns.

The performance levels in the consonant identification task (see Table 1) are substantially lower than in some previous studies, such as in Iverson et al. (1998), in which feedback was given to each response, and Walden et al. (1977), in which the participants trained before the test. The levels of performance in the present study provide a more ecologically valid estimate of how accurately phonemic information is perceived by persons with normal hearing. The present results most closely resemble those obtained by Owens and Blazek (1985) using /uCu/ stimuli, where 19% of the consonants were correctly identified by participants with normal hearing and 24% by participants with impaired hearing.

Word and sentence identification. For the word and sentence identification tasks, two separate repeated-measures factorial ANOVAs were performed. In word identification, there was a main effect of talker, $F(1, 23) = 138.84$, $p < .001$, $\eta^2 = .86$. In sentence identification, there were main effects of talker, $F(1, 23) = 48.07$, $p < .001$, $\eta^2 = .68$, and of topic, $F(1, 23) = 4.23$, $p < .05$, $\eta^2 = .16$. Thus, the human talker was more intelligible than the synthetic talker at both linguistic levels, and topical cues enhanced sentence identification performance. Further, word identification resulted in substantially higher scores than sentence identification, $t(23) = 10.32$, $p < .001$, $d = 2.37$, which replicates results of previous studies on word and sentence identification performance (e.g., Bernstein et al., 2000; Lidestam et al., 1999).

The combined results from consonant, word, and sentence identification reveal that there is a large and interesting difference in the data patterns for the human and the synthetic talker. There was no effect of talker on consonant identification, but there was a significant and large effect of talker on word and sentence identification. What is more, word and sentence identification scores were consistently higher than the consonant identification scores for the human talker, whereas they were consistently lower for the synthetic talker (see Tables 1 and 2). This suggests that either coarticulation is an important factor in perception of visual speech or that approximate consonant recognition is adequate for word and sentence recognition.

Consonant Identification: Sensitivity to the Phonemic Information

Sensitivity to single consonants. Sensitivity to the consonants was assessed with the d' metric (Green & Swets, 1966; Macmillan & Creelman, 1991) and computed with the DPrime Plus software (Creelman & Macmillan, 1996) by collapsing the confusion matrices and entering the frequencies of correct identification (i.e., hits), correct rejections, incorrect identification (i.e., false alarms), and failures of identification (i.e., misses) for

Table 3. Sensitivity (d') to consonants and confusions among consonants presented by the human talker.

	Response																		
Stimulus	**d**	**k**	**n**	**ŋ**	**r**	**j**	**g**	**ʈ**	**l**	**ɧ**	**s**	**t**	**h**	**v**	**f**	**b**	**m**	**p**	***d′***
d	8	10	20	9	4	22	6	13	10	13	4	12	3	3	4			3	.23
k	7	10	24	13	7	9	5	14	15	6	12	12	1	1	5	1	2		.39
n	5	6	14	11	13	7	3	6	55	3	1	5	9	1	1		2	1	.18
ŋ	7	5	12	12	8	24	8	7	31	6	6	8	5	1	1		1	1	.22
r	7	3	12	9	8	6	6	8	60	8	1	3	6		4	1	1	1	.14
j	12	7	22	5	9	21	6	8	6	12	6	6	13	4	3	1	1	1	.47
g	9	6	21	15	6	14	19	8	20	3	4	7	4	1	1	1	5		.75
ʈ	4	7	6	9	21	10	9	4	61	3	1	2	2			1	2	2	−.22
l		8	4	16	6	22	5	5	55	4	1	3	9	2		1	2	1	.87
ɧ	7	5	13	4	4	16	5	6	5	31	16	15	3	4	3		2	4	.96
s	3	6	10	10	1	6	5	13	5	21	38	17	4	1	1	1	1	1	1.17
t	9	1	15	10	8	8	3	6	3	11	30	30	4	2	1		3	2	.91
h	4	8	2	10	11	9	7	1	18	2	2	5	59	1	1	1	4		1.71
v	3	1	3	3				1	1	4				107	18		2	1	2.33
f	1	2	2	4	1	2	1	3		3	1	4	1	91	22	1	3	2	1.06
b	1		2	4		1	2	5	6	1	2			1	1	44	36	37	1.17
m	2		2	1	2		1	5			1			1	1	51	33	44	1.03
p	2		3	2			3	1				2				55	27	48	1.31
Total	91	85	187	147	109	177	94	114	351	131	126	131	123	221	67	159	127	149	

Note. The order of phonemes, from top to bottom, was the order of entry into the hierarchical cluster analysis.

each consonant. The columns to the far right in Tables 3 and 4 present the d' values. As can be seen in Tables 3 and 4, there was considerable variation in how accurately the consonants were perceived. Sensitivity values ranged between $d' = 2.33$ for /v/ spoken by the human talker to $d' = -.57$ for /ʈ/ spoken by the synthetic talker. Negative d' values reflect the fact that the stimuli had strong response biases. In other words, some stimuli were systematically identified as certain other stimuli.

For the synthetic talker, average d' across all consonants was .66. The average d' of .82 for the human talker across consonants indicates that sensitivity to consonants in visual speech perception (speechreading) under naturalistic circumstances is relatively low among individuals with normal hearing (i.e., below $d' = 1.0$). This was also reflected by the proportion of correctly identified consonants. The difference between talkers in average d' was nonsignificant ($p > .05$) when tested with a paired t test.

Visual similarity among consonants. Stimulus–response confusion matrices with collapsed responses from all participants in each of the two talker conditions are presented in Tables 3 and 4, respectively. Following the procedure of Iverson et al. (1998), these confusion data were transformed into phi-square values as similarity measures for all pairs of stimuli over the entire distributions of responses given to these stimuli.

Hierarchical cluster analysis was then performed on the symmetric matrices with the similarity (phi-square) values, using SPSS 10. An average-linkage-within-groups method was used for the clustering (see Iverson et al., 1998). Clusters of consonants were defined as visemes by Walden et al. (1977) if at least 75% of the responses were within-class, also replicating Iverson et al. Following this definition of Walden et al., three visemes were found for the human talker: /b m p/, /f v/, and /d k n ŋ r j g ʈ l/ (see Table 5). It should, however, be noted that the a priori probability of a response to either of the nine consonants in the /d k n ŋ r j g ʈ l/ cluster is 50%, which may be assumed to artificially inflate the within-cluster response rate as compared to the smaller clusters (cf. Walden et al.). For the synthetic talker, only the /p b m/ cluster satisfied the criterion of a viseme according to the criterion of Walden et al. (see Table 5).

The hierarchical clustering schemes for the human and the synthetic talker both consist of distinct /b m p/ and /f v/ clusters. This is in accordance with previous studies (e.g., Iverson et al., 1998; Walden et al., 1977). The /b m p/ and /f v/ clusters for both talkers also have in common that the individual consonants are perceptually distinct, as indicated by the relatively high d' values (see Tables 3 and 4). This is probably because their places of articulation (bilabial and labiodental) are more highly visible than for the other consonants.

Sensitivity to clusters of consonants. A d' was computed for each cluster of phonemes that emerged from the cluster analysis. This metric was used to get an appreciation of how successful individuals with normal

Table 4. Sensitivity (*d′*) to consonants and confusions among consonants presented by the synthetic talker.

	Response																		
Stimulus	**p**	**b**	**m**	**f**	**v**	**ŋ**	**k**	**r**	**ʃ**	**ʧ**	**d**	**s**	**t**	**n**	**j**	**g**	**h**	**l**	***d′***
p	67	36	15	1	3	4	2	2	2	4	1		3	2		2		1	1.33
b	70	37	10	1	1	2	7	3	1	5			1	2		2	1	1	1.18
m	78	23	25	2	3		2		1	3	3				1	1		2	.89
f	4	2	4	36	75	3		2		4	2	2	1			3	1	2	1.26
v	9	3	2	24	74	2	3	3	3	3	4		6	1	2	3		2	1.56
ŋ	4	3	6	3	7	16	6	12	3	11	15	2	3	15	3	8	9	18	.42
k		1	7	1	13	19	2	2	13		5	10	4	14	12	7	29	4	−.43
r	3		1	4	3	6	3	14	4	8	6	1	3	4	2	2	13	67	.31
ʃ				1	1	3		10	3	3		2		2		1		118	−.04
ʧ	1	1	7	6	13	11	7	11	8	1	3	9	2	9	12	7	26	5	−.57
d	4		4	3	5	7	11	15	1	4	16	4	8	13	7	19	4	18	.47
s	3	2	2	1	1	6	10	10	2	4	18	4	14	13	8	31	4	9	.09
t	5	1	4	3	3	4	5	10		5	32	3	8	14	9	21	5	9	.30
n	1	2	2	6	3	12	11	6	3	5	9	8	6	30	15	8	2	14	.86
j	1	4	8	3	7	12	14	11	1	2	5	2	4	17	25	16	5	5	.81
g	1	1		3	4	13	8	17	2	6	4	2	9	7	10	46	2	9	1.12
h		2	9	5	9	13	3	8	10	3	3	9		3	17	2	41	4	1.18
l	2				2	5	1	8	1	4	1	2	8		2	2	1	105	1.18
Total	253	118	106	103	227	138	95	144	58	75	127	60	80	146	125	181	143	393	

Note. The order of phonemes, from top to bottom, was the order of entry into the hierarchical cluster analysis.

hearing were at identifying consonants as members of the group of the most visually similar consonants. The procedure was the same as for single consonants, with the exception that the entries for clusters were based on hits, misses, correct rejections, and false alarms for the respective clusters of consonants. That is, hits were equal to the within-class response frequency.

Table 5 presents the within-class response rates, *d′* values, and 95% confidence intervals for the consonant clusters for both talkers. As can be seen in Table 5, the /b m p/ (bilabial) and /f v/ (labiodental) clusters yielded considerably higher *d′* values, compared with the other clusters, for both talkers. The confidence intervals for the *d′* values indicate that these differences are significant. Further, the remaining clusters (plus /h/) yielded *d′* values well above one for the human talker, but well below one for the synthetic talker. This suggests that for the the least visually prominent consonants, the human talker conveys subtle features that the synthetic talker fails to convey.

Table 5. Within-class response rate, sensitivity (*d′*), and 95% confidence intervals (CI 95%) for clusters of consonants by talker.

Cluster	Within-class response rate	*d′*	CI 95%
Human talker			
/b m p/	87%	3.05	±.19
/f v/	83%	2.96	±.21
/d k n ŋ r ʃ g ʧ l/	79%	1.48	±.11
/ʃ s t/	48%	1.34	±.14
/h/	41%	1.71	±.23
Synthetic talker			
/p b m/	83%	2.57	±.16
/f v/	73%	2.24	±.18
/d s t n ʃ g h l/	66%	.79	±.10
/ŋ k r ʃ ʧ/	25%	.23	±.12

There is an interesting discrepancy between the measures of perceptual accuracy. The proportion of correctly identified phonemes was at the same level for both talkers. The average *d′* values across consonants did not differ between the two talkers. However, mean word identification and sentence identification performance did differ significantly, and considerably in terms of effect size (see Lidestam et al., 2006). Sensitivity to clusters of consonants also differed, especially with regard to the large clusters of visually indistinct consonants. (When /h/ was included in the nearest cluster of the human talker, neither within-class response rate nor *d′* values were raised, and thus /h/ was excluded.) We suggest that sensitivity to the indistinct clusters is important for enabling meaningful percepts, since the indistinct visual phonemic information accounts for a major part of the visual speech signal. The more accurately a phoneme can be identified as belonging to a cluster, and the more specific (i.e., the smaller) this cluster is, the more constrained the lexical activation becomes, especially if there is linguistic

information from other, more well-defined phonemes in the same linguistic context. Under these circumstances, a lexical activation can produce a potent phonemic percept (Samuel, 1997). In other words, a small difference in the sensory and perceptual definition of phonemic information may be inflated when there is a linguistic context. Auer and Bernstein (1997) and Bernstein et al. (2000) showed that relatively small differences in perceived phonetic information can result in relatively large differences in accuracy in perception of words and sentences. Iverson et al. (1998) concluded that impoverished phonetic information may suffice for recognition of multisyllabic words, but that additional phonetic information may be needed for recognition of monosyllabic words. The results and conclusions of Bernstein and colleagues fit well with the hypothesis that sensitivity to clusters of the most impoverished phonetic information may play a more important role if other linguistic information has been perceived (e.g., if an individual has correctly identified enough phonemes). This hypothesis is also consistent with the cohort model of speech perception (Marslen-Wilson, 1987, 1989, 1995; Marslen-Wilson & Tyler, 1980; Marslen-Wilson, Moss, & van Halen, 1996). This model predicts that activation of the lexicon by previously identified phonemes facilitates the recognition of successive phonemes by activating cohorts of available words. Further, the model provides an alternative to the notion that coarticulation accounts for a key part of visually extractable phonetic information. However, further research is needed in order to pit these two hypotheses against each other.

Correlations

In 7 of 12 cases, consonant identification was significantly correlated ($p < .05$) with both word and sentence identification for the human talker. For the synthetic talker, only 1 of 12 instances was significant (see Table 6). This comparison of correlational patterns suggests that, in natural speech, the phonemic information is sufficiently specified to enable word (and, hence, sentence) identification for most individuals. However, the phonemic information conveyed by the synthetic talker may be too unspecified for most individuals to be able to identify words without topical information. The correlations may also reflect that in the synthetic talker conditions, the low performance levels in word and sentence identification performance entailed a small between-subjects variation, compared with the human talker conditions.

The correlations for the human talker, however, conform with previous studies (e.g., Bernstein et al., 2000; Demorest et al., 1996; Lidestam et al., 1999), such that the word and sentence identification tasks, when collapsed over conditions of topic and emotion, were moderately to highly correlated with each other ($r = .74$, $p < .01$). When word identification was collapsed over topic and emotion, the correlation was .60 ($p < .01$) with consonant identification; whereas the correlation between sentence and consonant identification was .67 ($p < .01$). The collapsed word and sentence identification scores were based on 36 items each, which allows better comparisons with the Bernstein et al. study, where 100 words and 25 sentences were spoken by the same male talker. For the synthetic face, there were no corresponding significant correlations between consonant identification, and word and sentence identification collapsed over topic and emotion. This may be because the phonemes are too ambiguous, especially with regard to how they form perceptual clusters, to enable word and sentence identification based on bottom-up processing, which requires sensitivity to the phonemes as building blocks of the signal.

Summary and Conclusions

The general purpose of this study was to investigate the role of visual perception of phonemes in speechreading among persons with normal hearing, that is, how the features of the signal are perceived and used together with complementary information, such as linguistic, paralinguistic, and topical cues.

The first specific purpose was to assess and describe how accurately phonemic information is perceived by

Table 6. Consonant identification correlations with word and sentence identification, respectively, by talker, topic, and emotion.

	Word identification						Sentence identification					
	Without topical cue			With topical cue			Without topical cue			With topical cue		
Talker	No emo.	Displ. emo.	CW emo.	No emo.	Displ. emo.	CW emo.	No emo.	Displ. emo.	CW emo.	No emo.	Displ. emo.	CW emo.
Human	.43*	.50*	.53**	.16	.26	.16	.45*	.39	.48*	.45*	.33	.41*
Synthetic	.28	.18	−.06	.11	.73**	.05	.10	−.12	−.13	.33	−.12	−.10

Note. No emo. = no cue to emotional valence of items; Displ. emo. = displayed emotion as cue to emotional valence of items; CW emo. = cue word to emotional valence of items.

*$p < .05$. **$p < .01$.

individuals with normal hearing under naturalistic conditions. The small proportion of correctly identified consonants and the generally low sensitivity values for single consonants suggest that most consonants usually are too visually indistinct to be identified correctly by individuals who have normal hearing. This was the case for the stimuli presented by both the human and the synthetic talker. Equal proportions of correctly identified consonants and about equal average sensitivity scores were observed for both talker conditions. Confusions among consonants resulted in clusters with varying degrees of cohesiveness (i.e., how early the clusters were joined in the analysis and how high the within-cluster response rate was). As a function of talker, the confusions among consonants resulted in varying degrees of visual distinctiveness of the clusters (i.e., how often the consonants of the cluster were identified as being one or other member of the cluster, as measured with d'). The most cohesive and distinct clusters for both talkers were the bilabial /b m p/ and labiodental /f v/ clusters. The remaining 13 consonants clustered differently and differed considerably with regard to how visually distinct they were depending on talker. That is, the clusters of the least visually distinct consonants, with the majority of the consonants, obtained considerably greater sensitivity values for the human talker than for the synthetic talker. This indicates that the ability to identify phonemes perfectly and the ability to identify phonemes adequately, as one of a restricted number of candidates, are not necessarily strongly associated.

The second specific purpose was to assess how consonant identification, where there is no semantic information but only a "pure" signal, is correlated with word and sentence identification, where top-down factors are inherent. Overall, consonant identification was moderately and positively correlated with both word and sentence identification for the human talker (cf. Bernstein et al., 2000). This can be compared to the relationship between topically cued speechreading performance and working memory capacity as measured with the reading span test (Rönnberg et al., 1989; after Baddeley, Logie, Nimmo-Smith, & Brereton, 1985). Lidestam et al. (1999) studied normal-hearing adults (n = 48) and found, after exclusion of 2 outliers, the correlation to be .58. Lyxell and Holmberg (2000) found a correlation of .49 for normal-hearing 11–14-year-olds. Thus, both phonemic sensitivity and some central aspect of working memory capacity appear to account for substantial parts of the variance in speechreading performance among normal-hearing individuals.

The final specific purpose was to explore what aspects of the phonemic information are most important to the perception of sentences and words. The present results suggest that the ability to identify a few single consonants is not sufficient for successful identification of words and sentences, whereas the additional ability to categorize less visually distinct consonants into small groups of potential candidates is highly useful. Considering how few consonants can be relatively easily identified as belonging to a distinct cluster, such as /b p m/ and /f v/, any additional signal features that can narrow down potential candidates for the remaining majority of consonants should be very useful. This finding is consistent with other studies, where additional information from another modality or a richer linguistic context has been found to aid identification of sentences and words (e.g., Auer & Bernstein, 1997; Iverson et al., 1998; Summerfield, 1983). This hypothesis can be tested experimentally by manipulating the parameters of consonant articulation in a synthetic talker. Further experiments comparing synthetic and natural visual speech will also benefit from including more than one human speaker, since human talkers may vary in terms of how much subtle phonemic information is conveyed.

Finally, clusters of phonemes could be defined using d' for the cluster, rather than within-cluster response rate (e.g., Walden et al., 1977). The d' measure taps identification rate as well as within-cluster response rate and error rate. We suggest $d' > 2$ for the definition of a meaningful viseme to in a multiple-choice identification task of at least 15 phonemes.

Acknowledgments

This study was in part funded by a grant from the Swedish Transport and Communication Research Board (1997-0603), awarded to Björn Lyxell.

We thank Björn Lyxell, Ulrich Olofsson, Jerker Rönnberg, Ulf Andersson, and Henrik Danielsson for comments on the manuscript; Mary Rudner for editing, help with phonetic transcriptions, and translation of stimuli into English; and all participants for their kind participation.

References

Andersson, U., & Lidestam, B. (2005). Bottom-up driven speechreading in a speechreading expert: The case of AA. *Ear & Hearing, 26,* 214–224.

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

Auer, E. T., Jr., & Bernstein, L. E. (1996). Lipreading supplemented by voice fundamental frequency: To what extent does the addition of voicing increase lexical uniqueness for the lipreader? In H. T. Bunnell & W. Idsardi (Eds.), *Proceedings ICSLP96, Fourth International Conference on Spoken Language Processing, Philadelphia* (pp. 86–89). Wilmington, DE: Applied Science and Engineering Laboratories.

Auer, E. T., Jr., & Bernstein, L. E. (1997). Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical

uniqueness. *Journal of the Acoustical Society of America, 102,* 3704–3710.

Auer, E. T., Jr., Bernstein, L. E., & Mattys, S. (2001). The influence of the lexicon on visual spoken word recognition. In D. W. Massaro, J. Light, & K. Geraci (Eds.), *Proceedings of the International Conference on Auditory-Visual Speech Processing* (pp. 7–12). Santa Cruz, CA: Perceptual Science Laboratory.

Baddeley, A. D., Logie, R., Nimmo-Smith, I., & Brereton, N. (1985). Components of fluent reading. *Journal of Memory and Language, 24,* 119–131.

Berger, K. W. (1972). Visemes and homophenous words. *Teacher of the Deaf, 70,* 396–399.

Bernstein, L. E., Demorest, M. E., & Eberhardt, S. P. (1994). A computational approach to analyzing sentential speech perception: Phoneme-to-phoneme stimulus–response alignment. *Journal of the Acoustical Society of America, 95,* 3617–3622.

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception & Psychophysics, 62,* 233–252.

Bernstein, L. E., Iverson, P., & Auer, E. T., Jr. (1997). Elucidating the complex relationships between phonetic perception and word recognition in audiovisual speech perception. In C. Benoît & R. Campbell (Eds.), *Proceedings of the ESCA–ESCOP Workshop on Audio-Visual Speech Processing* (pp. 89–92). Bonn, Germany: Institut für Kommunikationsforschung und Phonetik.

Beskow, J. (1995). Rule-based visual speech synthesis. In J. M. Pardo, E. Enríquez, J. Ortega, J. Ferreiros, J. Macías, & F. J. Valverde (Eds.), *Proceedings of Eurospeech '95, 1,* 299–302. Grenoble, France: ESCA.

Beskow, J. (1997). Animation of talking agents. In C. Benoît & R. Campbell (Eds.), *Proceedings of the ESCA–ESCOP Workshop on Audio-Visual Speech Processing* (pp. 149–152). Bonn, Germany: Institut für Kommunikationsforschung und Phonetik.

Carlson, R., Granström, B., & Hunnicutt, S. (1982). A multi-language text-to-speech module. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 7,* 1604–1607.

Creelman, C. D., & Macmillan, N. A. (1996). DPrime Plus [Computer software]. Retrieved July 14, 2006, from http://www.psych.utoronto.ca/~creelman/dprime5.zip

Demorest, M. E., Bernstein, L. E., & DeHaven, G. P. (1996). Generalizability of speechreading performance on nonsense syllables, words, and sentences: Subjects with normal hearing. *Journal of Speech and Hearing Research, 39,* 697–713.

Ellis, T., MacSweeney, M., Dodd, B., & Campbell, R. (2001). TAS: A new test of adult speechreading. Deaf people really can be better speechreaders. In D. W. Massaro, J. Light, & K. Geraci (Eds.), *Proceedings of the International Conference on Auditory-Visual Speech Processing* (pp. 13–17). Santa Cruz, CA: Perceptual Science Laboratory.

Green, D. M., & Swets, J. A. (1966). *Signal detection and psychophysics*. New York: Wiley.

Iverson, P., Bernstein, L. E., & Auer, E. T., Jr. (1998). Modeling the interaction of phonemic intelligibility and lexical structure in audiovisual word recognition. *Speech Communication, 26,* 45–63.

Lidestam, B. (2002). Effects of displayed emotion on attitude and impression formation in visual speech-reading. *Scandinavian Journal of Psychology, 43,* 261–268.

Lidestam, B., Beskow, J., & Lyxell, B. (2006). *Emotional cues in speechreading: Semantic cueing or enhanced articulatory specification?* Manuscript submitted for publication.

Lidestam, B., Lyxell, B., & Andersson, G. (1999). Speechreading: Cognitive predictors and displayed emotion. *Scandinavian Audiology, 28,* 211–217.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20,* 384–422.

Lyxell, B. (1994). Skilled speechreading: A single case study. *Scandinavian Journal of Psychology, 35,* 212–219.

Lyxell, B., & Holmberg, I. (2000). Visual speechreading and cognitive performance in hearing-impaired and normal hearing children (11–14 years). *British Journal of Educational Psychology, 70,* 505–518.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. New York: Cambridge University Press.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. *Cognition, 25,* 71–102.

Marslen-Wilson, W. D. (1989). Access and integration: Projecting sound into meaning. In W. D. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 3–24). Cambridge, MA: MIT Press.

Marslen-Wilson, W. D. (1995). Activation, competition, and frequency in lexical access. In G. Altmann (Ed.), *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 148–172). Cambridge, MA: Bradford.

Marslen-Wilson, W. D., Moss, H. E., & van Halen, S. (1996). Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 1376–1392.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition, 8,* 1–71.

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.

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

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science, 28,* 89–104.

Rönnberg, J. (1993). Cognitive characteristics of skilled tactiling: The case of GS. *European Journal of Cognitive Psychology, 5,* 19–33.

Rönnberg, J. (1995). What makes a skilled speechreader? In G. Plant & K.-E. Spens (Eds.), *Profound deafness and speech communication* (pp. 393–416). London: Whurr.

Rönnberg, J., Arlinger, S., Lyxell, B., & Kinnefors, C. (1989). Visual evoked potentials: Relation to adult speechreading and cognitive function. *Journal of Speech and Hearing Research, 32,* 725–735.

Rönnberg, J., Samuelsson, S., & Lyxell, B. (1998). Conceptual constraints in sentence-based lipreading in the hearing-impaired. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye I: Advances in the psychology of speechreading and auditory–visual speech* (pp. 143–153). Hove, England: Psychology Press.

Samuel, A. G. (1997). Lexical activation produces potent phonemic percepts. *Cognitive Psychology, 32,* 97–127.

Samuelsson, S., & Rönnberg, J. (1993). Implicit and explicit use of scripted constraints in lip-reading. *European Journal of Cognitive Psychology, 5,* 201–233.

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy: The principles and practice of numerical classification*. (2nd ed., rev.). San Francisco: Freeman.

Summerfield, Q. (1983). Audio-visual speech perception, lipreading, and artificial stimulation. In M. E. Lutman & M. P. Haggard (Eds.), *Hearing science and hearing disorders* (pp. 131–182). London: Academic Press.

van Son, N., Huiskamp, T. M. I., Bosman, A. J., & Smoorenburg, G. F. (1993). Viseme classifications of Dutch consonants and vowels. *Journal of the Acoustical Society of America, 96,* 1341–1355.

Walden, B. E., Prosek, R. A., Montgomery, A. A., Scherr, C. K., & Jones, C. J. (1977). Effects of training on the visual recognition of consonants. *Journal of Speech and Hearing Research, 20,* 130–145.

Young, S., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (1997). The HTK Book [Computer software manual]. Cambridge, England: Entropic Cambridge Research Laboratory.

Received September 21, 2004

Revision received April 18, 2005

Accepted January 21, 2006

DOI: 10.1044/1092-4388(2006/059)

Contact author: Björn Lidestam, Department of Behavioural Sciences, Linköping University SE-581 83 Linköping, Sweden. E-mail: bjli@ibv.liu.se

Appendix (p. 1 of 2). Examples of word and sentence identification items.

Words

Scenario: A Visit to a Restaurant

Positive Valence

öl [øːl] – beer
god [guːd] – good, tasty
nöjd [nœjd] – satisfied
mera [meːra] – more
toppen [tɔpen] – terrific
lyxig [lʏksɪg] – luxurious

Negative Valence

dyrt [dyːʈ] – expensive
beskt [beskt] – bitter
stank [staŋk] – stench
äcklig [ɛklɪg] – disgusting
långsam [lɔŋsam] – slow
avsmak [aːvsmɑːk] – distaste, disgust

Scenario: A Visit to the Doctor

Positive Valence

pigg [pɪg] – full of beans
bra [brɑː] – good, well
frisk [frɪsk] – well, healthy, fit
grattis [graːtɪs] – congratulations
bättre [bɛtre] – better
normal [nɔrmɑːl] – normal

Negative Valence

ont [ʊnt] – ache, pain, aching, painful
sjuk [ɧʉːk] – sick, ill
död [døːd] – dead, death
dålig [doːlɪg] – bad
tumör [tθmøːr] – tumor
smärta [smæʈa] – pain, ache

Appendix (p. 2 of 2). Examples of word and sentence identification items.

Sentences

Scenario: A Visit to a Restaurant

Positive Valence

Det var superbt! [dɛvɑ:sθpærbt] – It's superb.
Vad billig öl! [vɑbɪlɪgø:l] – The beer's a bargain.
Jättegod soppa! [jɛtəgu:dsɔpa] – Excellent soup.
Så här trevligt har jag inte haft på länge! [sohæ:rtre:vlɪthɑ:rjɑɪntəhaftpolɛŋə] – I haven't had such a nice time in ages.
Det här är verkligen ett prisvärt ställe! [dɛhærəværkɭigenetpri:svæ:ʈstɛlə] – This place is really good value for money.
Det var länge sedan jag åt något så gott! [dəvalɛŋəse:danjɑo:tno:gɔtso:gɔt] – It's a long time since I had such a delicious meal.

Negative Valence

Fruktansvärt dyrt! [frθktansvæ:dy:ʈ] – Daylight robbery.
Kom, vi går! [kɔmvɪgo:r] – Come on, we're leaving.
Inte gott alls! [ɪntəgɔtals] – Not very nice.
Nu har vi väntat i evigheter! [nəhɑ:rvɪvɛntatɪe:vɪghe:tər] – We've been waiting for ages.
Det här vill jag inte betala för! [dəhærvɪljɑ:ɪntəbəta:lafø:r] – I'm not paying for this.
Nu kallar jag på hovmästaren och klagar. [nɯ:kalarjɑ:poho:vmɛstarənɔklɑ:gar] – I'm going to complain to the manager.

Scenario: A Visit to the Doctor

Positive Valence

Det ser bra ut. [dəserbrɑ:ɯ:t] – It looks fine.
Du är frisk. [dθəfrɪsk] – You're fit.
Ingen fara! [ɪŋənfɑ:ra] – Don't worry.
Du behöver inte vara orolig. [dθbəhø:vərɪntəvarau:ru:lɪg] – There's nothing to worry about.
Vad roligt att du äntligen är frisk! [varu:lɪtatdθɛntlɪgenəfrɪsk] – How nice that you are well at last.
Operationen var mycket lyckad! [ɔpəraʃju:nənvɑ:mʏkəlʏkad] – The operation was very successful.

Negative Valence

Det ser hopplöst ut. [dəse:rhɔplø:stɯ:t] – The situation looks hopeless.
Jag är ledsen. [jɑɛlesən] – I'm sorry.
Du dör snart. [dθdø:ʂɳɑ:ʈ] – You will die soon.
Det kommer att göra ont. [dekɔməratjø:raʊnt] – It will hurt.
Du borde verkligen äta mindre. [dθbu:ɖəværkɭigenɛ:tamɪndrə] – You really should eat less.
Det finns nog inte så mycket mer att göra. [dəfɪnsnu:gɪntəsomʏkəme:ratjø:ra] – I'm afraid there's not much more we can do.

Visual Phonemic Ambiguity and Speechreading

Björn Lidestam, and Jonas Beskow
J Speech Lang Hear Res 2006;49;835-847
DOI: 10.1044/1092-4388(2006/059)