Helsinki University of Technology
Laboratory of Computational Engineering
Technical Report B47

Model-based assessment of factors influencing categorical audiovisual perception

Tobias S. Andersen

Dissertation for the degree of Doctor of Philosophy to be presented for public examination and debate in Auditorium S1 at Helsinki University of Technology on March 10th , 2005, at 12 o'clock noon.

Helsinki University of Technology
Department of Electrical and Communications Engineering
Laboratory of Computational Engineering

Teknillinen korkeakoulu
Sähkö- ja tietoliikennetekniikan osasto
Laskennallisen tekniikan laboratorio

Distribution:

Helsinki University of Technology
Laboratory of Computational Engineering
P. O. Box 9203
FIN-02015 HUT
FINLAND

Tel. +358-9-451 5325
Fax. +358-9-451 4830

E-mail: tobias@lce.hut.fi
http://www.lce.hut.fi
Online in PDF format: http://lib.hut.fi/Diss/

©Tobias S. Andersen

ISBN 951-22-7547-3 (printed)
ISBN 951-22-7548-1 (PDF)
ISSN 1455-0474

Picaset Oy
Helsinki 2005

Abstract

Information processing in the sensory modalities is not segregated but interacts strongly. The exact nature of this interaction is not known and might differ for different multisensory phenomena. Here, we investigate two cases of categorical audiovisual perception: speech perception and the perception of rapid flashes and beeps.

It is known that multisensory interactions in general depend on physical factors, such as information reliability and modality appropriateness, but it is not known how the effects occur. Here we parameterize the effect of information reliability for both our model phenomena. We also describe the effect of modality appropriateness as that of a factor that interacts with the effect of information reliability for counting rapid flashes and beeps.

Less explored is whether multisensory perception depends on cognitive factors such as attention. Here we show that visual spatial attention and attentional set influence audiovisual speech perception. Whereas visual spatial attention affected unimodal perception prior to audiovisual integration, attentional set influenced the audiovisual integration stage. We also show a strong effect of intermodal attention on counting rapid flashes and beeps.

Finally, we introduce a quantitative model, early maximum likelihood integration (MLI), of the interaction between counted flashes and counted beeps. We compare early MLI to the Fuzzy Logical Model of Perception (FLMP) which is a MLI model based on categorical percepts, and show that early MLI fits the data better using fewer parameters. Early MLI is also able to incorporate the effects of information reliability and intermodal attention in a more efficient way than the FLMP.

Author:
Tobias S. Andersen
M.Sc., Researcher
Laboratory of Computational Engineering
Helsinki University of Technology
Finland

Supervisors:
Kaisa Tiippana
PhD, Senior Lecturer
Laboratory of Computational Engineering
Helsinki University of Technology
Finland

Academy Professor Mikko Sams
Laboratory of Computational Engineering
Helsinki University of Technology
Finland

Preliminary examiners:
Professor Jari Hietanen
Department of Psychology
University of Tampere
Finland

Pentti Laurinen
Docent, PhD, Senior Lecturer
Department of Psychology
University of Helsinki
Finland

Official opponent:
Jean-Luc Schwartz
DR CNRS
Directeur de l'Institut de la Communication Parlée
UMR CNRS No 5009
Grenoble
France

List of Publications

The thesis consists of an overview of the following seven publications

P 1 Andersen, T.S., Tiippana, K. and Sams, M., Factors influencing audiovisual fission and fusion illusions, *Cognitive Brain Research*, 21 (2004) 301-308.

P 2 Tiippana, K., Andersen, T.S. and Sams, M., Visual attention modulates audiovisual speech perception, *European Journal of Cognitive Psychology*, 16 (2004) 457-472.

P 3 Andersen, T.S., Tiippana, K. and Sams, M., Visual Spatial Attention in Audiovisual Speech Perception, *Laboratory of Computational Engineering Technical Report B48. ISBN 951-22-7552-X. (2005).*

P 4 Tuomainen, J., Andersen, T.S., Tiippana, K. and Sams, M., Audio-visual speech perception is special, *Cognition, In Press* (2005).

P 5 Andersen, T.S., Tiippana, K. and Sams, M., Maximum Likelihood Integration of Rapid Flashes and Beeps*, Neuroscience Letters, In Press* (2005).

P 6 Andersen, T.S., Tiippana, K., Lampinen, J. and Sams, M., Modelling of Audiovisual Speech Perception in Noise, *Proceedings of AVSP 2001* (2001) 172-176.

P 7 Andersen, T.S., Tiippana, K. and Sams, M., Using the Fuzzy Logical Model of Perception in Measuring Integration of Audiovisual Speech in Humans, *Proceedings of NeuroFuzzy2002* (2002).

Description of the thesis author's contribution to Publications P1-P7

All of this work is a result of a joint effort between all the authors. The following reluctant extraction of my individual contribution is for thesis evaluation purposes.

In Publication P1, I was the principal author who conceived the idea of the study together with Kaisa Tiippana, conducted the experiments, performed the analysis and wrote the manuscript with the help of my co-authors at all stages.

In Publication P2, I contributed to the data analysis and the writing of the manuscript, and was responsible for the modeling section.

In Publication P3, I was the principal author. Mikko Sams conceived the original idea of Experiment 1. I conceived the idea of Experiment 2, conducted the experiments, performed the analysis and wrote the manuscript with the help of my co-authors at all stages.

In Publication P4, I contributed in conceiving the idea of the study, conducting the experiments, analyzing the data and writing the manuscript.

In Publication P5, I was the principal author who conceived the idea of the study, performed the analysis and wrote the manuscript with the help of my co-authors at all stages. This study was a theoretical study re-analyzing empirical work presented in Publication P1.

In Publication P6, I was the principal author who conceived the idea of the study, performed the analysis and wrote the manuscript with the help of my co-authors at all stages. This study was a theoretical study re-analyzing unpublished empirical work by Riikka Möttönen, Kaisa Tiippana and Mikko Sams.

In Publication P7, I was the principal author who conceived the idea of the study, performed the analysis and wrote the manuscript with the help of my co-authors at all stages. This study was a theoretical study re-analyzing empirical work presented in Publication P2.

Abbreviations

FLMP	Fuzzy Logical Model of Perception
EEG	Electro-encephalography
MLI	Maximum Likelihood Integration
MMN	Mismatch Negativity
SNR	Signal to noise ratio

Acknowledgements

The work described in this thesis was carried out in the Laboratory of Computational Engineering at Helsinki University of Technology under the supervision of Dr. Kaisa Tiippana and Academy Professor Mikko Sams. First and foremost, I wish to thank my supervisors. Kaisa Tiippana I thank for her tenacious work on revising our manuscripts and, hopefully, improving my writing skills in the process. In addition, she has been an invaluable help in all stages of my work. Mikko Sams has provided a relaxed, pleasant and supportive environment for academic work to flourish. He has also been of invaluable help particularly in the conception and completion of my work. The work on sine-wave speech was done in collaboration with Jyrki Tuomainen. I wish to thank him for asking me to join him in the project which turned out to be much more interesting than I would ever have imagined. I also wish to thank Aki Vehtari for fruitful discussions on modeling perception and for commenting on some of the manuscripts in this thesis. Thanks also to Professor Jouko Lampinen who was a catalyst in getting me started with modeling perception and provided some useful scripts for getting me started. The eye-movement measurements described in Publication P3 were made at Center for Innovation and Knowledge Research with Jari Laarni, Jaana Simola and Ilpo Kojo. I particularly wish to thank Jari Laarni who spent half a summer with me getting the device to work properly. In administrating the MUHCI project, my travel expenses and other bureaucratic matters, Eeva Lampinen has been a guardian angel without whom I am sure something terrible would have happened. Thanks, Eeva. Finally, thanks to everybody at the Laboratory of Computational Engineering who have made my daily life a good one for more than 4 years. The funding received from the EU Research Training Network MUHCI is gratefully acknowledged.

On a more personal note, I like to thank Linda Rosengren for entering my life and adding a little ‘golden rims’ to it. Anton Puolakka, Henrik Huhtinen, Marcus Engdahl, Miro Malmelin, Ove Holmqvist, Pami Aalto, Pasi Kurvinen, Peter Löfgren, Petrus Pennanen and the rest of the gang provided some very inspirational moments, warmth and immoral support along the way.

Table of Contents

Introduction

This work aims at understanding how categorical auditory and visual information is integrated in human perception. The focus has been on constructing a quantitative model of human responses with little regard to which particular neural system it is implemented. To understand a system, it is essential to understand how it is influenced by its surroundings—stimulus properties, in this case—and its state—cognitive factors, in this case. It has long been known that multisensory interactions in general depend on stimulus properties and the properties of the perceptual system but the role of cognitive factors is poorly understood. This work therefore puts the emphasis on those.

Audiovisual perception

In the past decades, it has become increasingly clear that information in the sensory modalities interacts strongly. From an information theoretical viewpoint this is not surprising. Using information from all available sources will generally yield better performance than using information from any one source of information and evolution will strive towards better performance. Therefore, evolution will strive towards integrating information from all the senses relevant for a given task. In this introduction, we shall focus on audiovisual integration, but will digress to other examples of multisensory integration that have been important in the history of the field. Perception can be divided into continuous and categorical perception. This distinction holds for audiovisual perception as well and will be described in the following.

Continuous perception

Continuous perception is when we perceive the variation of stimulus properties on a continuous scale—e.g. when we perceive location. When location varies continuously, so does our estimate of it. Location can be determined by either audition or vision. If we can both see and hear an object, the resulting location estimate is a compromise between the location estimated from audition or vision alone [2,56]. Other examples of continuous percepts that integrate across audition and vision are simultaneity [18,66] and intensity [55].

Categorical perception

Categorical perception is when continuously varying stimuli are perceived in categories. The full extent of categorical perception implies that the stimulus, or features of the stimulus, can be varied continuously but that this variation is only perceived when it crosses category boundaries. Two types of experiments are the standard requirement for determining whether a stimulus is perceived categorically. One experiment is an identification experiment. A stimulus attribute is varied continuously and subjects are to classify the stimulus. The transition between responding in one category and in the other should be abrupt for perception to be categorical. This transition is called the category boundary. The category boundary could, however, be due to an effect at the response level rather than at the perceptual level. Therefore, in the other required experiment, subjects are to discriminate between stimuli on the same continuum as used in the identification task. If subjects can only discriminate between stimuli separated by the category boundary, the effect is truly perceptual and perception is categorical; we perceive only the category of the stimulus, not continuous variations in its attributes.

Categorical perception was first demonstrated in speech perception by Liberman et al. [27]. However, it is important to note that this report did not find speech perception to be completely categorical. Rather, they found that discrimination is better across the category boundary, but that we do also have discrimination ability within categories. This weaker definition of categorical perception finds better support in the literature. It is also in accordance with neurophysiological findings. The mismatch negativity (MMN) is an electroencephalography (EEG) component that is elicited in response to infrequent auditory stimuli in a stream of frequent auditory stimuli. The difference between frequent and infrequent stimuli must be audible for the MMN to be elicited. For speech sounds, within-category differences elicit the MMN although between-category stimuli elicit a stronger MMN [39]. Yet another caveat in the theory of categorical perception is that whether a stimulus is perceived categorically or continuously may depend on the stimulus presentation and response scheme [20]. Here we shall assume a broad definition of categorical perception that merely implies that subjects respond in categories—not on a continuous scale.

One of example categorical audiovisual perception is the counting of rapid flashes and beeps. Shams and co-workers showed that one rapid flash may be perceived as two when accompanied by two rapid beeps [47,48]. The unaccompanied flash, however, is clearly perceived as a single flash. Thus this is an auditorily induced visual illusion. Shams et al. also showed that two rapid flashes are not perceptually fused into one when accompanied by a single beep. They generalized this phenomenon to stating that while there is a clear fission illusion (perceiving more flashes than actually presented due to more beeps) there is no fusion illusion (perceiving fewer flashes than actually presented due to fewer beeps). In Publication P1, we replicated the original work of Shams and co-workers [5] and, surprisingly, we found also a clear fusion illusion. In two later studies, Shams and coworkers demonstrated that electrophysiological correlates of the illusory flash occurred in visual cortex [10,49].

Speech perception affords a complex example of audiovisual integration and has accordingly been studied extensively. From speech, we perceive many qualities, both continuously and categorically, which can be derived from either face or voice. In accordance, many of them are based on information integrated across audition and vision when both sources of information are available. The continuous qualities include source location [8] and loudness [43] which have both been shown to be influenced by both audition and vision. The categorical include identity, gender, emotional tone and phonetic content. Of these, phonetic content [36] and emotional tone have been shown to integrate across audition and vision [19].

Phonetic speech perception stands out between all examples of multisensory integration due to the McGurk effect [36]. In McGurk and McDonald's classical example of this effect, a video of a face uttering /ga/ is dubbed with a voice saying /ba/. This results in an auditory percept of hearing /da/. If the utterances of the face and voice are interchanged so that a face uttering /ba/ is dubbed with a voice saying /ga/, the resulting auditory percept is /bga/. What makes the McGurk effect unique is that we have no comprehensive theory of how it arises from the auditory and visual percepts. Such theories exist for other

multisensory phenomena. The McGurk effect therefore stands unchallenged as the most complex and most poorly understood example of multisensory integration. Another effect of audiovisual integration of speech is that auditory speech comprehension is improved when watching a congruently articulating face [58]. This is, of course, the ecologically valid situation under which audiovisual integration of speech developed evolutionary.

Factors influencing categorical audiovisual perception

Physical factors

By physical factors we refer to the properties of the stimulus and stimulus transduction. Perhaps the first physical factor described in the literature on multisensory integration was modality appropriateness [71]. Some of the first studies on multisensory integration employed prism goggles that displace visual input [42,69]. Subjects were then given motor tasks like grasping [42] or pointing [69] where proprioceptive or tactile information was also available. This elicited the phenomenon of visual capture where the visual modality completely dominates or captures the other modality. Visual capture was explained by the modality appropriateness hypothesis: The more appropriate modality dominates perception. By more appropriate is understood having greater acuity for the task at hand. According to this hypothesis, vision dominates in spatial tasks because (primate) visual spatial acuity is greater than proprioceptive spatial acuity. Conversely, audition dominates in temporal tasks because temporal acuity is greater for audition than for vision. This is seen when the estimated rate of visual flicker is strongly influenced by the rate of concurrently presented auditory flutter [50,70].

However, later it was hypothesized that the basis of the more appropriate modality dominating perception was not its appropriateness *per se* but rather that it provided more reliable information due to its appropriateness. This is termed the information reliability hypothesis. It predicts that if the more appropriate modality receives less reliable information it looses its advantage and will no longer dominate.

Unfortunately, information reliability has been used in two meanings in the literature. Whereas Schwartz used the term to describe the stimulus' reliability [46], Warren used it

to denote the perceived reliability which includes subjects' assumptions on its reliability which he manipulated by telling naive subjects about the diffracting properties of the prism goggles they wore [68]. To distinguish between these two meanings, we adopt the terms stimulus and cognitive information reliability. Certainly, the two concepts are related in that both physical and cognitive information reliability may contribute to, what we shall call perceptual information reliability. In this section, we shall describe physical stimulus information reliability and, in the next section on cognitive factors, we shall briefly revisit cognitive information reliability.

Stimulus information reliability denotes how informative a stimulus about an attribute of an object [71]. The most straightforward example of a stimulus attribute that affects information reliability is the stimulus signal-to-noise ratio. A noisy stimulus—i.e. a stimulus with a low signal-to-noise ratio (SNR)—is less reliable, or informative. However, information reliability may not always depend so simply on stimulus SNR; the reliability of certain speech features are more robust to variations in SNR than others [37].

On the basis of the information reliability hypothesis, some of the phenomena that formed the basis of the modality appropriateness hypothesis have been revisited by researchers varying stimulus information reliability to investigate whether information reliability or modality appropriateness is the governing principle of multisensory interactions. Ernst and Banks studied the effect of information reliability on haptics and concluded that "Visual dominance occurs when the variance associated with visual estimation is lower than that associated with haptic estimation" [16]. Wada et al. studied the effect of information reliability on the perceived auditory flutter and visual flicker rates and concluded that "When ambiguous auditory temporal cues were presented, the change in the frequency of the visual stimuli was associated with a perceived change in the frequency of the auditory stimuli" [67]. Thus, in these types of experiment that originally led to the modality appropriateness hypothesis, the more reliable, not the more appropriate modality, dominate perception.

The information reliability hypothesis does not abolish the modality appropriateness hypothesis. It is not stimulus information reliability that determines modality dominance in multisensory perception. Rather, it is the perceptual reliability. The perceptual reliability is determined by both the stimulus reliability and the modality appropriateness, or acuity. Perceptual information reliability cannot exceed modality acuity regardless of the stimulus information reliability.

How do the principles of information reliability and modality appropriateness apply to audiovisual categorical perception? In Publication P1, we present the only study so far that has formally investigated these effects on audiovisual integration of rapid flashes and beeps [5]. This study consisted of two experiments. In both Experiments, the stimuli consisted of 1-3 flashes, 1-3 beeps and all possible audiovisual combinations of flashes and beeps. In Experiment 1, the beeps were at a clearly audible sound level and in Experiment 2, the beeps were near detection threshold. Each experiment consisted of two blocks. In the count-flashes block the visual and audiovisual stimuli were presented and subjects were instructed to count the flashes. In the count-beeps block the auditory and audiovisual stimuli were presented and subjects were instructed to count the beeps. We found that the number of concurrent beeps strongly influenced the number of perceived flashes. The effect was stronger when the beeps were at a clearly audible level than when they were near subjects' auditory threshold, which is in accordance with the information reliability hypothesis. We also found a converse, visually induced auditory illusion were the number of concurrently presented flashes influenced the perceived number of beeps but only when the beeps where near detection threshold. This also supports the information reliability hypothesis. Notably, the effect of information reliability was much stronger on the visually induced auditory illusion, which completely disappeared just above auditory threshold, than on the auditorily induced visual illusion, which persisted near auditory threshold even with clearly visible flashes. This effect can be explained by the modality appropriateness hypotheses. Audition influenced vision throughout a greater range of stimulus information reliability than vision influenced audition because audition had greater acuity for the task of counting rapid events.

The effect of information reliability and modality appropriateness on audiovisual speech perception is somewhat more complicated. Obviously, audition is the more appropriate modality for speech comprehension, but the McGurk effect shows that vision exerts a strong influence on audition even at clearly audible sound levels. One explanation for this comes from the manner-place hypothesis [29]. There are three main features characterizing consonants: Voicing, manner and place. By voicing is meant that the voice is activated during a consonant uttering. Voiced consonants are e.g. /g/, /b/ and /d/. Their unvoiced counterparts are /k/, /p/ and /t/ respectively. Voicing is a purely auditory speech cue. There is no visual difference between a face uttering /k/ and /g/. In accordance with the modality appropriateness hypothesis, voicing should not be influenced by visual speech at all, which is indeed the case [29]. However, place of articulation is often more pronounced in vision than in audition. As an example, take the similarly sounding consonants /m/ and /n/ where visually the bilabial closing of the /m/ is clearly distinguishable from the alveolar /n/. Thus, for certain speech cues, vision may actually be the more appropriate modality. Auditory and visual speech thus complement each other and this might be the basis of why speech perception is integrated across audition and vision.

Several studies have shown visual influence on the auditory speech percept increases with decreasing SNR [30,58] which is in accordance with the information reliability hypothesis. In Publication P6 we present a study where we confirmed this result using the McGurk illusion [3]. As an example, our study showed that a clear McGurk effect of hearing /apta/ when the auditory stimulus was /ata/ and the visual stimulus was /apa/ gradually changed into hearing /apa/ when the auditory SNR, and thus auditory information reliability was lowered, showing a stronger visual influence of audition.

We conclude that the hypotheses of information reliability and modality appropriateness, which were first shown to apply to multisensory continuous perception, also apply to audiovisual categorical perception.

Cognitive factors

Most studies in the literature have concluded that audiovisual integration of speech is an automatic process that does not depend on the perceiver's cognitive state [15,31,36,53,65]. This is surprising seen in the light of Treisman's feature integration theory of attention [61]. According to this theory, attention is necessary for features to be assembled into objects. This applies to feature integration across as well as within sensory modalities [60]. However, Liberman has argued that speech perception occurs in a specialized module that functions automatically without the need for attention [28].

In the literature, there is a tendency to form a dichotomy between perceptual and cognitive effects, which we find is not completely justified [19,48]. Certainly, there is an important difference between response bias and perceptual effect. In the Stroop effect, the word "red" colored blue infers a response bias towards "red" when subjects are to identify the color of the word [57]. There is no perceptual effect of actually seeing the blue word as red, so this is a response bias. The McGurk effect, however, is truly a perceptual effect as witnessed by those who have experienced it [36]. This is supported by neurophysiological studies that show that visual speech can modify activity in parts of the brain associated with conscious auditory perception [38,40,44]. So far, the dichotomy is justified. But perceptual effects *can* certainly depend on attention. A recent striking example is provided by studies on inattentional blindness [51]. Simons and Chabris showed that a clearly visible person dressed up as a gorilla for conspicuity might not be consciously perceived when subjects are performing a moderately demanding visual task concurrently [52]. Here, the dichotomy between perceptual and cognitive effects is unjustified because the effect is truly perceptual but depends on attention. Therefore, audiovisual integration of speech *can* depend on attention even though it is a truly perceptual effect.

Here, we shall first review the evidence for audiovisual integration of speech being an automatic process. Then, we shall review our own work which has shown that this is not always the case.

In the first report on the McGurk effect, McGurk and McDonald noticed that the illusion persists after extended exposure even when subjects are aware of the discrepancy between the face and the voice [36]. This certainly indicates that the illusion and hence audiovisual integration is not under voluntary control under the circumstances employed.

Massaro has studied the effect of instructing subjects to report either what they heard or what they saw on perception of phonetic and emotional contents of speech [31,32]. He found a strong effect of task instructions in that visual influence was greater when subjects reported what they saw and that auditory influence was greater when they reported what they heard. Since the stimuli were the same in both cases, this could indicate an effect of task instructions on audiovisual integration. However, Massaro reached another conclusion that the effect occurred at the unimodal processing stage. His conclusion was based on a model-based analysis, to which we shall return in the section on models of audiovisual integration.

Massaro's conclusion has found acceptance in the literature. Driver accepted Massaro's argument as a confirmation of his own conclusion that the audiovisual integration of speech source location seems to be automatic and uninfluenced by cognitive factors [15]. Driver studied the interaction of auditory spatial attention and audiovisual integration of spatial location. He conducted an experiment using a loudspeaker on each side of a display of visual speech. Speech coming from one loudspeaker was congruent with the visual speech; speech coming from the other was not. Seeing the visual speech impaired participants' ability to identify the incongruent speech token. Driver concluded that the visual speech integrated with the congruent auditory speech causing an illusory displacement of the auditory sound source towards the central display and hence the other sound source. The illusory decrease in distance between the two sound sources impeded the participants ability to separate one from the other and hence comprehension. The audiovisual integration was thus detrimental to task performance, and on this basis Driver concluded that audiovisual integration of talker location occurs involuntarily and automatically.

Another study pointing to automatic audiovisual integration was done by Soto-Faraco who studied a syllabic interference task [53]. Two syllables were presented in rapid consecution. The second syllable only interferes with the classification of the first if it varies; not if it is constant. This is interpreted as auditory attention failing to select the relevant stimulus so that the second syllable is obligatorily processed even though participants attempt to focus their attention only on the first one. By using audiovisual incongruent stimuli as the second syllable they created a McGurk effect. It was the illusory percept rather than the actually presented acoustic stimulus that determined whether interference occurred. Soto-Faraco et al. concluded that the visual influence on the auditory speech percept must have occurred before attentional selection.

Vroomen et al. studied the effect of performing a secondary task while perceiving the emotional content of a face and voice [65]. In order to scan for an effect of attention across attentional faculties, three secondary tasks were employed in separate experiments; addition, counting and pitch discrimination. In all three experiments did a static face influence the perceived emotional content of the voice similarly.

The above studies show that audiovisual integration of some speech attributes is robust to some cognitive factors. The coverage of speech attributes and cognitive factors is, however, sparse and we did not find them sufficient conclude by induction that audiovisual integration of speech is generally robust to variations in *all* cognitive factors. Therefore, we proceeded with the following studies of the effect of attention on audiovisual speech perception.

In Publication P2, we employed a leaf like drawing appearing to float across a talking face [59]. Subjects were instructed to follow the leaf with their gaze while reporting what they heard the talker say. At the moment of articulation the leaf was near the face so that gaze displacement would have little effect on speech perception—visual or audiovisual. Still, we did find a decrement in visual influence on auditory speech perception in that the McGurk illusion was weaker when subjects attended the leaf. This decrement must then have been due to the attentional demands of the tracking task including the displacement

of visual spatial attention. The decrement was, importantly, not always accompanied by a decrement in lip reading performance in the corresponding unimodal visual condition. This points to an effect on audiovisual integration rather than an effect on unimodal perception prior to integration.

In Publication P3, we continued to study the effect of visual spatial attention [7] using visual stimuli consisting of two faces on each side of a central fixation point. Below the fixation point was an arrow cueing the subjects which face to attend. In the visual condition, subjects were instructed to lipread the attended face. In the audiovisual condition, a voice was dubbed onto the movie of the faces and subjects were instructed to respond according to what they heard. In the conclusive experiment in this study one face uttered /eke/, the other /ete/ while the voice uttered /epe/. This type of stimulus enabled us to distinguish the effect on perception that each face and the voice had.

We found a strong effect of visual spatial attention in that the attended face had a greater influence on speech perception than did the unattended face. The unattended, distractor face did however also influence speech perception indicating that attentional selection was not complete. This was seen in the control condition employing videos without the distractor face—i.e. with only one face laterally displaced from the central fixation point. Here, the effect of the face on speech perception was greater than when the distractor face was present. The influence of the voice and thus audiovisual integration was unaffected by the distractor face. Assuming that the attentional load was greater when the distractor face was present, this indicates that attention worked independently of audiovisual integration. Therefore, the effect of visual spatial attention is likely to have taken effect prior to audiovisual integration at the unimodal visual processing level. This was corroborated by the results from unimodal visual trials in which the effects of visual spatial attention and the distractor face were similar to the results in audiovisual trials.

While both our studies on the effect of visual spatial attention on audiovisual speech perception show a significant effect, they disagree on whether this effect of attention occurs at the unimodal processing stage or at the audiovisual integration stage. In the first

study, the effect of attention on audiovisual perception was not reflected in unimodal visual trials indicating that the effect occurred at the audiovisual integration stage. In the second study, the effect of attention was reflected in unimodal visual trials indicating that the effect occurred at the unimodal visual stage. However, in the last study we found that the effect of the distractor face was reflected in unimodal perception only when subjects' eye movements were monitored. This could be due to subjects looking – despite instructed not to - at the attended object in the visual but not audiovisual condition only when they were not aware of being monitored. The reason why this would occur in the visual but not audiovisual condition could be that in the visual condition they reported on the attended face which would be easier if they also looked at it. In the audiovisual condition they reported on the voice which would not get any easier if they looked at the attended face. Unfortunately, this difference in motivation was also present in the first study. Since we did not record eye movements in that study, we cannot check whether absence of an effect of attention in the unimodal visual condition was due to eye movements. The second, more controlled study therefore forms the basis of our conclusion.

To summarize, our studies on the effect of visual spatial attention on audiovisual speech perception indicate that there is a strong effect. The effect occurs prior to and independent of audiovisual integration so that visual spatial attention selects the face to be integrated with the auditory speech percept. This agrees well with the studies outlined above that also found no effect of attention on audiovisual integration. However, most of these studies showed no effect of attention on audiovisual integration by showing no effect of attention on audiovisual perception. In contrast, our studies showed no effect of attention on audiovisual integration *despite* a strong effect of attention on audiovisual perception. We have thus demonstrated that the effects of attention at the unimodal processing stage can propagate to bimodal perception, and that it is important to distinguish between the effects of attention on audiovisual perception and audiovisual integration.

In Publication P4, we found an effect of attention which is likely to target audiovisual integration per se [62]. In this study we employed sine-wave speech created by positioning time-varying sine waves at the centre frequencies of the three lowest formants of a natural speech signal. Two speech tokens, /omso/ and /omso/ were used. Our pilot studies showed that when the natural speech token /omso/ was dubbed onto a face saying /onso/ a McGurk effect of hearing /onso/ occurred. Likewise /onso/ dubbed onto a face saying /omso/ was heard as /omso/.

First, subjects were trained to categorize the sine-wave speech tokens in two arbitrary categories. At this point, the subjects were not aware of the speech like nature of the stimuli. We then tested that they were able to perform this task. Then we used audiovisual stimuli consisting of the sine-wave speech tokens dubbed onto the face uttering the speech tokens. Of special interest are the incongruent audiovisual speech stimuli where we would expect to see a McGurk effect similar to that for natural speech. If the McGurk effect occurred, concurrent presentation of the incongruent face should change subjects' responses to the other response category. This effect was, however, very weak. We then trained subjects in perceiving the sine-wave speech tokens as speech and to categorize them phonetically. Then we tested the same audiovisual speech stimuli as before and now found a very strong McGurk effect. We interpreted this result as evidence of a speech specific mode of audiovisual perception which can be manipulated by cognitive factors. Only when subject were aware of the speech like nature of the sine-wave speech stimuli did they enter speech mode and integrated sine-wave speech with visual speech.

Finally, we return to our study of audiovisual integration of rapid flashes and beeps in Publication P1 [5]. Our results showed a strong effect of task instructions, or intermodal attention, on subjects' responses. When counting flashes, subjects' responses were strongly influenced by the number of concurrently presented clearly audible beeps. When counting clearly audible beeps, subjects' responses were uninfluenced by the number of concurrently presented flashes. This shows that instructing subjects to respond according to what they see increases influence from vision while instructing subjects to respond

according to what they hear increases influence from audition. This effect was quite similar to that found by Massaro on audiovisual speech perception described above [31,32].

Recall the distinction between physical and cognitive information reliability from above. Cognitive information reliability denotes the effect of our assumption of the information reliability [68]. It is easy to see that intermodal attention can be seen as assuming that the unattended modality is less reliable. Thus the effect of attention could be to increase the perceptual information reliability of the attended modality. This is in accordance with the gain theory of attention stating that the effect of attention is to increase the gain of the attended stimulus relative to the unattended stimulus through increased processing [25,63].

In summary, our studies on the effect of attention on audiovisual speech perception showed that there is a great effect of visual spatial attention which occurs prior to audiovisual integration at the unimodal processing level. Further, our study showing an effect of intermodal attention on counting of rapid flashes and beeps extends Massaro's finding of an effect of intermodal attention on audiovisual speech perception to an effect on categorical audiovisual perception in general. This effect shall be analyzed in greater detail below. Finally, the effect of speech mode on sine-wave speech perception strongly indicates that cognitive factors can influence the audiovisual integration stage of speech perception.

Models of categorical audiovisual perception

A multitude of functional architectures for audiovisual integration in categorical perception has been suggested [46]. We find it most informative to divide models in two categories: early and late. Early models assume that audiovisual integration occurs prior to classification. Features extracted from the auditory and visual speech signals are integrated and the integral forms the basis for a categorical decision. Late models assume that classification occurs in each modality and that these classifications are then integrated across audition and vision. We describe these two classes of models in the following.

Early integration

Although not quantitative, the manner-place hypothesis described above has had some success in describing which McGurk illusion arises from which audiovisual stimulus combinations and why [29]. The manner-place hypothesis is an early model; it is manner and place that are integrated audiovisually before phonetic classification. Its success should serve as an encouragement for deriving a quantitative, or exact, early model. In Publication P5, we introduced a model called early maximum likelihood integration (MLI) [6] which we applied to the perception of rapid flashes and beeps. Here, we describe this model first because other early models can be derived as from it.

Maximum likelihood as the principle governing multisensory integration has recently been studied for stimuli falling on a continuum [2,16,17]. In these studies, it has been assumed that the stimulus, *S*, causes an internal representation, *x*, in the brain. In the process, perceptual Gaussian noise is added so that the probability of an internal representation value given a stimulus is given by:

$$P(x \mid S) = \sqrt{\frac{r}{2\pi}} \exp\left(-\frac{r(x-\mu_S)^2}{2} \right) \tag{1}$$

where μ_S and r denote mean and reliability of the internal representation, respectively. The reliability, r, relates to the standard deviation, σ, of the Gaussian distribution as

$$r = \frac{1}{\sigma^2} \tag{2}$$

Given an auditory stimulus, S_{A}, and a visual stimulus, S_{V}, with internal representation means μ_A and μ_V respectively, the integrated internal representation, x_{AV}, is also Gaussian distributed with mean

$$\mu_{\mathrm{AV}} = w\mu_{\mathrm{A}} + (1-w)\mu_{\mathrm{V}} \tag{3}$$

where the weight, w, is

$$w = \frac{r_A}{r_A + r_V} \tag{4}$$

and the reliability is

$$r_{AV} = r_A + r_V \tag{5}$$

It is reasonable to equate the reliability in this model with perceptual information reliability as described above. Then, this model contains the information reliability hypothesis in the form of Eq. 4. When a modality is more reliable it is weighted higher.

In order to apply this model to categorical responses, it is necessary to add a model of categorization. We applied a simple model based on signal detection theory [24]. A category, C, is defined by an interval, $[x^C_{\min}, x^C_{\max}]$ of internal representation values. One endpoint may be replaced with plus or minus infinity as appropriate. When a feature value falls inside the interval, the stimulus is estimated to belong to the category; when it falls outside the interval, the stimulus is estimated not to belong to the category. The probability of a stimulus being classified as belonging to category C is then

$$\begin{aligned} &P(C \mid S) = \\ &P\left(x^C_{\min} < x < x^C_{\max} \mid S\right) = \\ &\int_{x^C_{\min}}^{x^C_{\max}} \frac{1}{\sigma_M \sqrt{2\pi}} \exp\left(\frac{(x-\mu_S)^2}{2\sigma_M^2}\right) dx = \\ &\Phi\left(\frac{\mu_S - x^C_{\min}}{\sigma_M}\right) - \Phi\left(\frac{\mu_S - x^C_{\max}}{\sigma_M}\right) \end{aligned} \tag{6}$$

where Φ is the standard normal probability function.

With this model of categorization, early MLI is a complete model applicable to audiovisual integration of a single feature. For multiple features, the internal representation is vectorial and not scalar as above. Then, the simple classification model no longer holds but needs to be extended.

This model has several desirable properties which come directly from the maximum likelihood rule for continuously perceived stimuli. First, it assigns system noise as the cause of response variability within subjects. This is in accordance with the well established signal detection theory [24]. Second, the weighting factor provides a measure of the relative influence of the sensory modalities involved. This is a model ability that has been sought after in the literature and which finds practical applications [21,33]. Third, it inherently incorporates the information reliability hypothesis which we have seen to be central in multisensory perception.

A similar model has been proposed by Braida [11] who called it a *pre-labeling model* referring to audiovisual integration preceding phonetic classification or labeling. This model did not apply the simplifying assumption that a single feature characterizes both auditory and visual stimuli, but that a single feature characterizes auditory stimuli and another single and independent feature characterizes visual stimuli. Braida and co-workers could thus test it on limited cases of audiovisual speech perception [11,12]. However, in order to reduce the complexity of the model, they assumed that the reliabilities of both modalities were the same. As we have seen, the information reliability and manner-place hypotheses indicate that this is not a good assumption. They also assumed that each modality carries only one feature and that this feature is not the same in each modality. For the features they used, voicing in the auditory modality and place in the visual modality, this is a rough assumption as place is mediated by both audition and vision. Still, Braida et al. obtained promising results in applying their model to a number of data sets and later Grant et al. obtained similar results [21,23].

Berthommier also applied a similar model, which he called Articulatory Feature Coding to spectrally reduced auditory speech [9]. Speech was spectrally reduced because certain

features of the manner-place hypothesis are localized in certain sub-bands of the auditory spectrum. This enabled Berthommier to vary auditory speech features in isolation which simplifies the model. However, Berthommier also assumed that the reliabilities of the modalities were identical. Still, also his results were promising but, to our knowledge, they have not been followed by a more comprehensive study.

Late integration

Our knowledge on audiovisual categorical perception and the factors affecting it provides restraints for quantitative models. The most striking effect is that of illusory percepts from incongruent audiovisual stimuli. This effect is closely linked to the increased classification performance to congruent audiovisual stimuli as compared to unimodal stimuli. Any model must be able to describe these phenomena. As Massaro has reviewed [31], there exists a plethora of late integration models of audiovisual speech perception that do not meet this basic requirement. We shall not include these models in this review. We shall include only Massaro's influential Fuzzy Logical Model of Perception (FLMP) and disregard a number of models which have been shown to be equivalent, or very similar, to it [13,35].

The FLMP has been suggested by Massaro to be a universal law for integrating information from multiple sources in cognitive systems [31,35]. Applied to audiovisual integration, the formula for the FLMP is

$$P(R_i \mid A,V) = \frac{P(R_i \mid A) \times P(R_i \mid V)}{\sum_{j=1}^{N} P(R_j \mid A) \times P(R_j \mid V)} \tag{7}$$

Here, $P(R_i \mid A,V)$ denotes the probability of responding in the ith response category given auditory, A, and visual, V, stimuli. The FLMP is recognized maximum likelihood rule of integrating two discrete, or categorical, independent probability distributions.

The FLMP has been tested by Massaro mainly on audiovisual speech perception but also on a multitude of other phenomena [31,32]. Generally, the model has fitted the data

extremely well. This fact, along with the theoretical attractiveness of a maximum likelihood model for multisensory integration has given it recognition in the literature. It has, however, also been the object of severe criticism, which we shall describe in the following.

In Publication P7, we showed that the FLMP is highly unstable when auditory and visual speech stimuli are incongruent—i.e. the stimulus combinations that can give rise to the McGurk effect [4]. Let us assume that the two sources of information A and V disagree about a binomial response probability, so that $P(R_1|A)=p_1$ and $P(R_1|V)=1-p_1$. This symmetrically incongruent information leads to a combined response probability of 50% which is obtained by inserting unimodal response probabilities into the FLMP:

$$P(R_1 \mid A,V) = \frac{p_1(1-p_1)}{p_1(1-p_1)+(1-p_1)p_1} = \frac{1}{2} \tag{8}$$

If the unimodal response probabilities change so that $P(R_1|A)=p_2=2p_1$ but $P(R_1|V)$ remains the same $P(R_1|V)=1-p_1$. Then the FLMP yields the combined response probability

$$P(R_1 \mid A,V) = \frac{2p_1(1-p_1)}{2p_1(1-p_1)+(1-2p_1)p_1} \approx \frac{2}{3} \tag{9}$$

The approximation holds when $p_1 \ll 1$. Thus, if a unimodal response probability is close to zero, it can often, with negligible effect on the goodness-of-fit, be chosen so that an *arbitrarily* small change in it can accommodate almost any accompanying change in the bimodal response probability. A similar argument as this was proposed by Schwartz [45] who called it the *zero-zero trick*.

This instability of the FLMP means that it could only be implemented in a noise-free system. But, the brain is certainly not noise-free, so the instability poses a great challenge for the FLMP. One way to meet this challenge would be to allow variations in the

unimodal response probabilities by assuming them to be distributed according to a Dirichlet-multinomial distribution rather than a multinomial distribution [26]. In the Dirichlet-multinomial distribution, the unimodal response probabilities are themselves distributed according to a Dirichlet distribution—i.e. they are noisy and vary randomly from trial to trial. We are not aware of any closed-form derivation of the maximum likelihood rule for integration two independent Dirichlet-multinomial distributions, but it is certainly not equivalent to the FLMP. Since this model would incorporate the effect of noise it would be stable when implemented in a real, noisy system.

The FLMP is a Generalized Linear Model known as the *baseline logit* model [1]. The instability of the FLMP has only been shown for incongruent auditory and visual information. This is reflected in the literature on Generalized Linear Models where it is emphasized that iterative maximization of likelihood is not a valid method for sparse categorical data. By sparse data is meant with many empty response categories, which is exactly the case when auditory and visual information is highly incongruent.

Instability as that shown by the FLMP is symptomatic for ill-posed problems. Problems are ill-posed when the number of degrees of freedom exceeds the number of data points. Note, that it is not the number of data points that determines ill-posedness. If two data points co-vary they share the same degree of freedom and should be considered as a single effectual data point. Also, if a data point does not show variation then it reflects no degree of freedom—i.e. we do not gain any information by adding empty response categories. Therefore, General Linear Models, such as the FLMP, can become ill-posed when the data is sparse. With more degrees of freedom than data points, there is no unique solution to ill-posed problems. Certainly, we can apply an iterative error-minimizing algorithm to obtain a parameter estimate, but this estimate is likely not to be unique and to depend on the initial conditions. Schwartz has demonstrated that FLMP solutions obtained with iterative algorithms might not be unique [45]. The cure against ill-posedness is regularization which is a smoothing of the solution. Using the Dirichlet-multinomial distribution rather than the multinomial has been suggested as a regularization method for models for categorical data [54]. This is in good agreement

with the notion that the Dirichlet-multinomial maximum likelihood model would be stable to noise. Furthermore, the goodness-of-fit does not represent how well ill-posed models describe the underlying structure of the system of interest. Optimization algorithms will use the excess free parameters to capture the coincidental structure in the sampling noise. This is called *over-fitting*. But, since there is not reproducible structure in the noise, fits to ill-posed models will not be able to predict new data even though they might describe the data well in retrospect. This criticism has also been directed towards the FLMP [14,41,64]. However, Massaro has answered the criticism by employing model evaluation criteria such as cross-validation [31] and Bayes factor [34] which take model-flexibility into account. Studies by other authors have also pointed to some predictive ability of the FLMP [11,22]. The topic of whether the FLMP is generally over-fitting remains controversial in the literature and, we believe, will remain so in the absence of a more parsimonious model.

If our concepts are fuzzy in the way described by the FLMP, then a response rule is necessary for transforming the fuzzy concept into a categorical response. The optimal response rule is the maximum a posteori rule which states that the response should always fall in the most likely category. However, the FLMP applies the equal probability rule which states that the responses are given probabilistically which is a suboptimal strategy. Thus the FLMP presents us with a black box perceptual system that reaches unimodal noiseless, but fuzzy percepts, which are then integrated according to an optimal integration rule only to be followed by a suboptimal response rule. All in all, this is a scenario which is difficult to align with our general understanding of perception and probability.

In the FLMP, perceptual information reliability is not explicit. The FLMP does however obey the information reliability hypothesis. If a modality is less reliable, the response probability distribution has higher entropy—i.e. it is closer to a uniform random distribution. If a modality is more reliable, it has lower entropy—i.e. it is closer to a peaked distribution. A totally unreliable modality causing a uniform random, flat response distribution will not influence bimodal perception according to the FLMP. A

totally reliable modality causing a response distribution with a 100% of responses in one category will completely dominate bimodal perception according to the FLMP. But this effect is not parameterized, so to model e.g. a change in auditory SNR it is necessary to re-estimate the auditory response probability distribution which requires N-1 degrees of freedom where N is the number of response categories. If the effect was parameterized a change of parameter value would suffice costing only 1 degree of freedom making the model more parsimonious.

Therefore the FLMP is in need of a parameterization of information reliability. In Publication P6, we suggested a model [3] which describes the auditory response probability distribution at a certain SNR as a weighted sum of a totally reliable and a totally unreliable response probability distribution.

$$P(R_i \mid A, SNR) = (1-\alpha)P(R_i \mid A) + \alpha/N \tag{10}$$

Here, $P(R_i \mid A, SNR)$ is the probability of responding in the ith response category given auditory stimulus, A, at a given SNR. The parameterization of the noise level (inverse reliability) is denoted α. This model can be inserted into the FLMP as to provide this model:

$$P(R_i \mid A, V, SNR) = \frac{\left[(1-\alpha)P(R_i \mid A) + \alpha / N\right] \times P(R_i \mid V)}{\sum_{j=1}^{N}\left[(1-\alpha)P(R_j \mid A) + \alpha / N\right] \times P(R_j \mid V)} \tag{11}$$

We tested this model on responses from an auditory, visual and audiovisual speech identification task where the auditory SNR was varied between 4 levels. For this experimental paradigm the parameterization of the auditory SNR reduced the number of free parameters by 18. Although it gave a poorer fit even when the error measure was corrected for the number of degrees of freedom, it did capture some of the salient features of the effect of varying the auditory SNR on perception. This bears some promise that more advanced model of the effect of information reliability could successfully be merged into the FLMP to obtain a satisfactory parameterization.

Comparative model testing

Early MLI has so far only been developed for audiovisual integration of one feature. This severe shortcoming hinders it being tested on audiovisual speech perception, which is the phenomenon that the FLMP has been applied to most often. However, counting rapid flashes and beeps is likely to be based on a single feature, so for this phenomenon both models apply. In Publication P5 [6], we conducted a study where we used the data from our study on counting rapid flashes and beeps in Publication P1 [5] to test and compare early MLI and the FLMP. Recall that in our study we varied auditory SNR between a clearly audible level and a near threshold level. We also varied task instructions between counting flashes and counting beeps. When we fitted the models separately to these four conditions, we found that early MLI fitted the data slightly better across all subjects and conditions. This should be seen in the light of early MLI having 38 fewer free parameters across all subjects and conditions. This result strongly favors early MLI.

Incorporating the effect of attention in maximum likelihood models

Maximum likelihood models of multisensory integration are stimulus driven, or bottom-up. They do not incorporate cognitive or top-down effects. At first, this seems to be in discrepancy with the two effects of attention on audiovisual categorical perception that we have described above: intermodal attention and attentional set, but a closer look will reveal that this is not necessarily the case. Here, we shall first examine the effect of intermodal attention which has caused some difficulties in the literature. We point to and correct a flaw in previous FLMP based analyses of the effect. We then briefly discuss the effect of attentional set, or speech mode, which is too recent to have been discussed in any but the original report, Publication P4.

Recall that Massaro's studies showed that instructing subjects to respond according to what they see increased visual influence compared to instructing subjects to respond according to what they hear. Massaro analyzed this effect using the FLMP [31,32]. To test the FLMP when subjects responded according to what they saw, he fitted the FLMP to the data from three conditions: the audiovisual condition when attending vision (AV/V) and unimodal auditory (A) and visual (V) conditions. To test the FLMP when

subjects responded according to what they heard, he again fitted the it to the data from three conditions: the audiovisual condition when attending audition (AV/A) and unimodal auditory (A) and visual (V) conditions. Massaro found no significant difference in FLMP goodness-of-fit between the two fits. He concluded that the FLMP described integration in both cases so there was no difference in audiovisual integration between the two conditions and that the effect therefore occurred before audiovisual integration at the unimodal processing stage.

In Publication 6 [6], we pointed to a paradox in Massaro's analysis. If we explicitly note the obvious fact that the unimodal auditory condition (A) was, in fact, the auditory condition when audition was attended (A/A), and likewise that the visual condition (V) was actually the visual condition when vision was attended (V/V) then we see that Massaro fitted both (AV/V) and (AV/A) to (A/A) and (V/V). The attentional state was thus varied only in the audiovisual condition but not in the unimodal conditions. We claim that (AV/A) should be fitted to (A/A) and (V/A), i.e. the visual condition when *audition* was attended. Likewise (AV/V) should be fitted to (V/V) and (A/V), i.e. the auditory condition when *vision* was attended. In that way, the attentional state would not vary within fits. Only thus would we isolate the audiovisual integration mechanism from the effect of intermodal attention. We emphasize that this problem applies generally to models of multisensory integration, but, to our knowledge, has never been recognized before. Of course, the (A/V) and (V/A) conditions are impossible. Subjects cannot respond to stimuli in one modality while attending another modality in which no stimulus occurs. These hypothetical conditions should therefore be left as free parameters. We note that the ability of the FLMP to provide good fits when applied in a paradoxical way reflects its purported hyper-flexibility. It is a remarkable expression of non-linear flexibility that the FLMP can describe widely different audiovisual percepts from identical unimodal percepts.

As described above, we studied the effect of instructing subjects to attend either vision or audition on the perception of rapid flashes and beeps and found an effect similar to that found by Massaro [6]. We fitted the FLMP to these data in the same way as Massaro

fitted it to his data. We also fitted early MLI to our data. As a maximum likelihood rule, early MLI faces the same problem as the FLMP in accounting for the effect of task instructions. Rather than using the Massaro's approach, we assumed that the effect of intermodal attention could be quantified as an effect upon perceptual information reliability so that attending a modality would lead to an increase in the perceptual reliability of that modality as compared to when it is not attended. As described earlier, this assumption is based on the gain theory of attention which equates the effect of attending a stimulus as an increase in the gain of the internal representation of that stimulus relative to the unattended stimulus or background noise [25,63]. Notably, due to the inherent parameterization of information reliability in early MLI, modeling the effect of intermodal attention on unimodal perception imposed the added cost of only one free parameter. Auditory perception when vision was attended (A/V) was modeled by decreasing the reliability of auditory perception when audition was attended (A/A). This approach is not only more reasonable but also much more parsimonious employing only 12 free parameters to model two attentional states where the FLMP employs 24 free parameters. In addition to these benefits, early MLI fitted the data better than the FLMP.

No maximum likelihood rule—neither early MLI nor the FLMP—provide any solution to the binding problem. There is no parameter that quantifies the amount of integration in isolation from unimodal perception. However, early integration might provide an explanation of how sine wave speech can be more or less bound to visual speech depending on whether it is perceived as speech. When sine wave speech is not perceived as speech, attention could focus on some acoustic feature that is irrelevant to phonetic classification. This feature might not be mediated by vision. Audition would therefore be the only reliable modality for this feature and classification would not be influenced by vision. When sine wave speech is perceived as speech, attention would focus on features relevant for phonetic classification. This feature would be likely to be mediated also by vision which therefore would influence perception.

Concluding summary

We have demonstrated how the information reliability hypothesis can incorporate many of the phenomena we see in multisensory perception. The stimulus information reliability

combined with modality acuity, or appropriateness, combines to form perceptual information reliability, which determines the relative influences of the sensory modalities in multisensory perception.

We then introduced early maximum likelihood integration (MLI) as a model for audiovisual categorical perception. This model is an extension of MLI for continuously perceived stimuli. Both models inherently and explicitly incorporate the information reliability hypothesis. Using the gain theory of attention, we showed that effects of attention can be incorporated in early MLI through changes in perceptual information reliability. We compared this model with the Fuzzy Logical Model of Perception (FLMP) and found that early MLI can describe data better while being a more parsimonious model. We conclude that the early MLI is a promising new model of audiovisual categorical perception in terms of goodness-of-fit, parsimony and interpretability.

Finally, we have shown that sine-wave speech is only integrated with visual speech when subjects are aware its speech-like nature which suggests that there may be other factors that influence audiovisual integration beyond the scope of current MLI models.

References

[1] Agresti, A., *Categorical Data Analysis*, John Wiley and Sons, Hoboken, New Jersey, 2002.

[2] Alais, D. and Burr, D., The ventriloquist effect results from near-optimal bimodal integration, *Curr Biol*, 14 (2004) 257-62.

[3] Andersen, T.S., Tiippana, K., Lampinen, J. and Sams, M., Modelling of Audiovisual Speech Perception in Noise. *Proceedings of AVSP 2001*, Ålborg, Denmark, 2001, pp. 172-176.

[4] Andersen, T.S., Tiippana, K. and Sams, M., Using the Fuzzy Logical Model of Perception in Measuring Integration of Audiovisual Speech in Humans, *Proceedings of NF2002* (2002).

[5] Andersen, T.S., Tiippana, K. and Sams, M., Factors influencing audiovisual fission and fusion illusions, *Cognitive Brain Research*, 21 (2004) 301-308.

[6] Andersen, T.S., Tiippana, K. and Sams, M., Maximum Likelihood Integration of Rapid Flashes and Beeps, *Neuroscience Letters, In Press* (2005).

[7] Andersen, T.S., Tiippana, K. and Sams, M., Visual Spatial Attention in Audiovisual Speech Perception, *Submitted* (2004).

[8] Bertelson, P., Vroomen, J., Wiegeraad, G. and de Gelder, B., Exploring the Relation Between McGurk Interference and Ventriloquism. *Proceedings of the International Congress on Spoken Language Processing*, Yokohama, Japan, 1994.

[9] Berthommier, F., Audio-visual recognition of spectrally reduced speech. *Proceedings of AVSP 2001*, Ålborg, Denmark, 2001.

[10] Bhattacharya, J., Shams, L. and Shimojo, S., Sound-induced illusory flash perception: role of gamma band responses, *Neuroreport*, 13 (2002) 1727-30.

[11] Braida, L.D., Crossmodal integration in the identification of consonant segments, *Q J Exp Psychol A*, 43 (1991) 647-77.

[12] Braida, L.D., Sekiyama, K. and Dix, A.K., Integration of audiovisually compatible and incompatible consonants in identification experiments. *Auditory-Visual Speech Processing (AVSP)*, Terrigal - Sydney, Australia, 1998.

[13] Cohen, M.M. and Massaro, D., On the similarity of categorization models. In F.G. Ashby (Ed.), *Multidimensional models of perception and cognition*, Erlbaum, Hillsdale, NJ, 1992, pp. 395-447.

[14] Cutting, J.E., Bruno, N., Brady, N.P. and Moore, C., Selectivity, scope, and simplicity of models: a lesson from fitting judgments of perceived depth, *J Exp Psychol Gen*, 121 (1992) 364-81.

[15] Driver, J., Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading, *Nature*, 381 (1996) 66-8.

[16] Ernst, M.O. and Banks, M.S., Humans integrate visual and haptic information in a statistically optimal fashion, *Nature*, 415 (2002) 429-33.

[17] Ernst, M.O. and Bulthoff, H.H., Merging the senses into a robust percept, *Trends Cogn Sci*, 8 (2004) 162-169.

[18] Fendrich, R. and Corballis, P.M., The temporal cross-capture of audition and vision, *Percept Psychophys*, 63 (2001) 719-25.

[19] de Gelder, B., Vroomen, J. and Pourtois, G., Multisensory Perception of Emotion, Its Time Course, and Its Neural Basis. In G. Calvert, C. Spence and B.E. Stein (Eds.), *The Handbook of Multisensory Processes*, MIT Press, Cambridge, Massachusetts, 2004, pp. 581-596.

[20] Gerrits, E. and Schouten, M.E., Categorical perception depends on the discrimination task, *Percept Psychophys*, 66 (2004) 363-76.

[21] Grant, K.W. and Seitz, P.F., Measures of auditory-visual integration in nonsense syllables and sentences, *J. Acoust. Soc. Am.*, 104 (1998) 2438-2450.

[22] Grant, K.W. and Walden, B.E., Predicting auditory-visual speech recognition in hearing-impaired listeners. *Proceedings of the XIIIth International Congres of Phonetic Sciences*, *Vol. 3*, 1995, pp. 122-129.

[23] Grant, K.W., Walden, B.E. and Seitz, P.F., Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration, *J. Acoust. Soc. Am.*, 103 (1998) 2677-2690.

[24] Green, D. and Swets, J., *Signal Detection Theory and Psychophysics*, John Wiley and Sons, Inc., New York, 1966, 455 pp.

[25] Hillyard, S.A., Mangun, G.R., Woldorff, M.G. and Luck, S.J., Neural Systems Mediating Selective Attention. In M.S. Gazzaniga (Ed.), *The Cognitive Neurosciences*, MIT Press, Cambridge, Mass., 1995.

[26] Johnson, N.L., Kotz, S. and Balakrishnan, N., *Discrete Multivariate Distributions*, John Wiley & Sons, New York, 1997.

[27] Liberman, A.M., Harris, K., Hoffman, H.S. and Griffith, B., The discrimination of speech sounds within and across phoneme boundaries, *Journal of Experimental Psychology*, 54 (1957) 358-368.

[28] Liberman, A.M. and Mattingly, I.G., The motor theory of speech perception revised, *Cognition*, 21 (1985) 1-36.

[29] MacDonald, J. and McGurk, H., Visual influences on speech perception processes, *Percept Psychophys*, 24 (1978) 253-257.

[30] MacLeod, A. and Summerfield, Q., Quantifying the contribution of vision to speech perception in noise, *British Journal of Audiology*, 21 (1987) 131-141.

[31] Massaro, D., *Perceiving talking faces*, The MIT Press, Cambridge, 1998, 495 pp.

[32] Massaro, D.W., *Speech Perception by ear and eye: A paradigm for psychological inquiry*, Erlbaum, Hillsdale, NJ, 1987.

[33] Massaro, D.W. and Cohen, M.M., Tests of auditory-visual integration efficiency within the framework of the fuzzy logical model of perception, *J Acoust Soc Am*, 108 (2000) 784-9.

[34] Massaro, D.W., Cohen, M.M., Campbell, C.S. and Rodriguez, T., Bayes factor of model selection validates FLMP, *Psychon Bull Rev*, 8 (2001) 1-17.

[35] Massaro, D.W. and Friedman, D., Models of integration given multiple sources of information, *Psychol Rev*, 97 (1990) 225-52.

[36] McGurk, H. and MacDonald, J., Hearing lips and seeing voices, *Nature*, 264 (1976) 746-748.

[37] Miller, G.A. and Nicely, P.E., An analysis of perceptual confusions among some English consonants, *The Journal of the Acoustical Society of America*, 27 (1955) 338-352.

[38] Möttönen, R., Krause, C.M., Tiippana, K. and Sams, M., Processing of changes in visual speech in the human auditory cortex, *Brain Res Cogn Brain Res*, 13 (2002) 417-25.
[39] Näätänen, R., The Perception of Speech Sounds by the Human Brain as Reflected by the Mismatch Negativity (MMN) and its Magnetic Equivalent (MMNm), *Psychophysiology*, 38 (2001) 1-21.
[40] Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I.P., Möttönen, R., Tarkiainen, A. and Sams, M., Primary auditory activation by visual speech: an fMRI study at 3 Tesla, *Neuroreport*, In Press (2004).
[41] Pitt, M.A., Data fitting and detection theory: reply to Massaro and Oden (1995), *J Exp Psychol Learn Mem Cogn*, 21 (1995) 1065-7.
[42] Rock, I. and Victor, J., Vision and Touch: An Experimentally Created Conflict between the Two Senses, *Science*, 143 (1964) 594-6.
[43] Rosenblum, L.D. and Fowler, C.A., Audiovisual investigation of the loudness-effort effect for speech and nonspeech events, *J Exp Psychol Hum Percept Perform*, 17 (1991) 976-85.
[44] Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O.V., Lu, S.T. and Simola, J., Seeing speech: visual information from lip movements modifies activity in the human auditory cortex, *Neurosci Lett*, 127 (1991) 141-5.
[45] Schwartz, J.-L., Why the FLMP should not be appliced to McGurk data: Or how to better compare models in the Bayesian framework. In J.-L. Schwartz, F. Berthommier, M.A. Cathiard and D. Sodoyer (Eds.), *Audio Visual Speech Processing (AVSP 2003), ISCA Tutorial and Research Workshop on Audio Visual Speech Processing*, St. Jorioz, France, 2003, pp. 77-82.
[46] Schwartz, J.-L., Robert-Ribes, J. and Escudier, P., Ten years after Summerfield: a taxonomy of models for audio-visual fusion in speech perception. In R. Campbell, B. Dodd and D. Burnham (Eds.), *Hearing by Eye II: Advances in the Psychology of Speechreading and Auditory-visual Speech*, Psychology Press, Hove, U.K., 1998, pp. 85-108.
[47] Shams, L., Kamitani, Y. and Shimojo, S., Illusions. What you see is what you hear, *Nature*, 408 (2000) 788.
[48] Shams, L., Kamitani, Y. and Shimojo, S., Visual illusion induced by sound, *Brain Res Cogn Brain Res*, 14 (2002) 147-52.
[49] Shams, L., Kamitani, Y., Thompson, S. and Shimojo, S., Sound alters visual evoked potentials in humans, *Neuroreport*, 12 (2001) 3849-52.
[50] Shipley, T., Auditory Flutter-Driving of Visual Flicker, *Science*, 145 (1964) 1328-30.
[51] Simons, D.J., Attentional capture and inattentional blindness, *Trends Cogn Sci*, 4 (2000) 147-155.
[52] Simons, D.J. and Chabris, C.F., Gorillas in our midst: sustained inattentional blindness for dynamic events, *Perception*, 28 (1999) 1059-74.
[53] Soto-Faraco, S., Navarra, J. and Alsius, A., Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task, *Cognition*, 92 (2004) B13-23.

[54] Steck, H. and Jaakkola, T., On the Dirichlet prior and Bayesian regularization. *Advances in Neural Information Processing, Vol. 15*, MIT Press, Massachusetts, 2002.
[55] Stein, B.E., London, N., Wilkinson, L., K. and Price, D., D., Enhancement of Perceived Visual Intensity by Auditory Stimuli: A Psychophysical Analysis, *Journal of Cognitive Neuroscience*, 8 (1996) 497-506.
[56] Stein, B.E. and Meredith, M.A., *The Merging of the Senses*, A Bradford Book, Cambridge, MA, 1993, 211 pp.
[57] Stroop, J.R., Studies of Interference in Serial Verbal Reactions, *Journal of Experimental Psychology*, 12 (1935).
[58] Sumby, W.H. and Pollack, I., Visual contribution to speech intelligibility in noise, *The Journal of the Acoustical Society of America*, 26 (1954) 212-215.
[59] Tiippana, K., Andersen, T.S. and Sams, M., Visual attention modulates audiovisual speech perception, *European Journal of Cognitive Psychology*, 16 (2004) 457-472.
[60] Treisman, A., Properties, parts and objects. In K.R. Boff, L. Kaufman and J.P. Thomas (Eds.), *Handbook of Perception and Human Performance. Cognitive Processes and Performance, Vol. 2*, John Wiley & Sons, New York, 1986, pp. 35.31-35.62.
[61] Treisman, A. and Gelade, G., A feature integration theory of attention, *Cognitive Psychology*, 12 (1980) 97-136.
[62] Tuomainen, J., Andersen, T.S., Tiippana, K. and Sams, M., Audio-visual speech perception is special, *Cognition, In Press* (2005).
[63] Verghese, P., Visual search and attention: a signal detection theory approach, *Neuron*, 31 (2001) 523-35.
[64] Vroomen, I.I. and Gelder, B., Crossmodal integration: a good fit is no criterion, *Trends Cogn Sci*, 4 (2000) 37-38.
[65] Vroomen, J., Driver, J. and de Gelder, B., Is cross-modal integration of emotional expressions independent of attentional resources?, *Cogn Affect Behav Neurosci*, 1 (2001) 382-7.
[66] Vroomen, J. and de Gelder, B., Temporal ventriloquism: sound modulates the flash-lag effect, *J Exp Psychol Hum Percept Perform*, 30 (2004) 513-8.
[67] Wada, Y., Kitagawa, N. and Noguchi, K., Audio-visual integration in temporal perception, *Int J Psychophysiol*, 50 (2003) 117-24.
[68] Warren, D.H., Spatial localization under conflict conditions: is there a single explanation?, *Perception*, 8 (1979) 323-37.
[69] Warren, D.H. and Cleaves, W.T., Visual-proprioceptive interaction under large amounts of conflict, *J Exp Psychol*, 90 (1971) 206-14.
[70] Welch, R.B., DuttonHurt, L.D. and Warren, D.H., Contributions of audition and vision to temporal rate perception, *Percept Psychophys*, 39 (1986) 294-300.
[71] Welch, R.B. and Warren, D.H., Immediate perceptual response to intersensory discrepancy, *Psychol Bull*, 88 (1980) 638-67.