Jekosch, U. (**2005**). “Assigning meaning to sounds – semiotics in the context of product-sound design” in: J. Blauert (ed.): *Communication Acoustics*, pp. 193–219, Springer, Berlin–Heidelberg–New York NY.

Juslin, N. P. and Västfjäll, D. (**2008**). “Emotional responses to music: The need to consider underlying mechanisms” Behav. and Brain Sciences **31**, 559–621

Lenneberg, H. E. (**1967**). *Biologische Grundlagen der Sprache* (biological foundations of language). Suhrkamp, D-Frankfurt/Main.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert–Kennedy, M. (**1967**). “Perception of speech code”. Psychol. Rev. **74**, 431-461.

Libermann, A. M., and Mattingly, I. G. **(1989)** “A specialization for speech perception” Science **243**, 489–494.

Lungwitz, H. (**1925**). *Die Entdeckung der Seele – Allgemeine Psychobiologie* (the discovery of the psyche − general psychobiology). De Gruyter, D–Berlin (cited here: 5th edition, De Gruyter, D–Berlin, **1947**).

Lungwitz, H. (**1933**). *Die Psychobiologie der Sprache* (the psychobiology of speech/language). Brückeverlag Kurt Schmersow, D–Kirchhain (cited here: 3rd, revised edition, Becker, R., ed., Thieme, Stuttgart–New York NY, **2010**).

Machleidt, W., Gutjahr, L., and Mügge, A: (**1989**). *Grundgefühle* (basic feelings). Springer, Berlin–Heidelberg–New York NY.

Maturana, H. R.: (**1987**). “Biology of language: the epistemology of reality” In: Miller, G.A., and Lenneberg. E. (eds.), *Psychology and biology of language and thought,* pp. 27−63. Academic Press, New York NY.

Mees, U. (**1985**). “What do we mean when we speak of feelings? On the psychological texture of words denoting emotions” Sprache und Kognition **1**, 2–20.

Ortony, A. and Clore, G. L. (**1989**). “Emotions, moods and conscious awareness”. Cogn. & Emotion **3**, 125–137.

Ortony, A. and Turner, T. J. (**1990**). “What’s basic about basic emotions” Psych. Rev. **97**, 315–331.

Panksepp, J. (**1982**). “Towards a general psychobiological theory of emotions”. Behav. and brains sciences. **5**, 407–467.

Plutchik, R. (**1962**). *The emotions: facts, theories, and a new model.* Random House, New York NY.

Rizzolatti, G. and Craighero, L. (**2004**). “The mirror-neuron system”. Ann. Rev. Neuroscience **27**, 169–192

Audiovisual integration in speech perception: a multi-stage process

KASPER ESKELUND[1] , JYRKI TUOMAINEN[2] AND TOBIAS ANDERSEN[1]

[1] *Cognitive Systems, Department of Informatics and Mathematical Modelling, Technical University of Denmark, DK-2800 Lyngby, Denmark*

[2] *Speech Hearing and Language Sciences, University College London, UK*

Integration of speech signals from ear and eye is a well-known feature of speech perception. This is evidenced by the McGurk illusion in which visual speech alters auditory speech perception and by the advantage observed in auditory speech detection when a visual signal is present. Here we investigate whether the integration of auditory and visual speech observed in these two audiovisual integration effects are specific traits of speech perception. We further ask whether audiovisual integration is undertaken in a single processing stage or multiple processing stages.

INTRODUCTION

Integration effects such as the McGurk effect (McGurk and MacDonald, 1976) and the detection advantage associated with audiovisual speech (Grant and Seitz, 2000) show that vision and hearing are integrated in speech perception. It is, however, unknown whether the processes underlying such audiovisual integration are specific for perception of speech, or if they pertain to audiovisual perception in general. Moreover, audiovisual integration is often tacitly assumed to be undertaken in a single step (Massaro, 1998; Vatakis and Spence, 2007). In the experiment reported here, we test whether audiovisual integration as seen in the McGurk effect and the audiovisual detection advantage occurs for both non-speech and speech perception. We further test these integration effects as to investigate whether they show different properties in non-speech and speech conditions. If the latter is the case, it may indicate that the effects are related to dissociated processes supporting the claim that audiovisual integration of speech is multi-faceted.

Grant and Seitz (2000) showed that seeing a synchronous visual speech signal is advantageous when detecting an acoustic speech signal masked by noise. Presenting three sentences in audiovisual and auditory-only formats masked by acoustic noise, they found that the advantage associated with the presence of the visual speech signal in the audiovisual stimulus was equivalent to a 1.6 dB gain of the auditory-only stimulus. Investigating the dynamics of the acoustic and visual stimuli, they showed that the magnitude of the advantage depends on the degree of correlation between changes in lip opening area and sound intensity. On this basis, they proposed the *peak listening hypothesis*, stating that cues in the visual signal guides

Proceedings of ISAAR 2011: Speech perception and auditory disorders. 3rd International Symposium on Auditory and Audiological Research. August 2011, Nyborg, Denmark. Edited by T. Dau, M. L. Jepsen, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN 87-990013-3-0. EAN 9788799001330. The Danavox Jubilee Foundation, 2012.

the listener to the spectral and temporal parts of the acoustic signal with the most favourable signal-to-noise ratio.

Experimenting with single-syllable audiovisual speech stimuli, Bernstein and colleagues (2004) found that preparatory lip gestures preceding acoustic onset may be responsible for the effect. Thus, even if the visual stimulus was exchanged with a non-speech geometric figure, it still evoked a detection advantage as long as the onset of the preparatory articulatory movements was retained. The authors concluded that the effect was not specific for speech stimuli and could be produced by any visual pre-cueing of auditory onset.

In a similar experiment, however, Schwartz and colleagues (Schwartz *et al*., 2004) observed that the audiovisual detection advantage was eliminated for non-speech visual stimuli, even if the dynamics of speech was represented. Further, in the results of Bernstein and co-workers (2004), the detection advantage was lower for non-speech visual stimuli. However, it is difficult to determine whether these findings are due to the geometrical visual stimuli lacking relevant cues present in natural visual speech or whether they are due to observers not using available cues because the non-speech figures seem irrelevant to the observer and visual cues are thus to a lesser degree bound together with the auditory signal. These studies have thus targeted a contrast between visual stimuli in addition to the difference between non-speech and speech perception, or, the perceptual set. Thus, the findings on the speech-specificity of the audiovisual detection advantage are inconclusive. In contrast, the purpose of the current experiment was to ask directly if the stimulus needs to represent speech.

As any stimulus containing a minimum of phonetic cues will be perceived as speech, it is difficult to devise a meaningful comparison of speech perception with non-speech perception using the same stimulus. Tuomainen *et al*. (2005) provided an elegant solution to this, using sine wave speech (SWS) stimuli (Remez *et al*., 1981). In SWS, centre frequencies of the three lowest formants of a natural speech token are extracted. A novel stimulus is generated by letting three sinusoids track these frequencies and their amplitudes. This synthetic stimulus thus contains only faint phonetic cues. When listening to SWS, naïve subjects tend not to perceive any phonetic content, but rather report hearing synthetic, meaningless sounds. However, when informed on the phonetic content, the weak phonetic cues are perceived and SWS heard as speech. Remez and colleagues (1981) interpreted this as evidence for a speech-specific mode of perception. Since SWS can be perceived as speech or as non-speech it is an ideal stimulus for investigating effects that supposedly occur specifically in speech perception.

With this approach, Tuomainen *et al*. (2005) found that the McGurk illusion only occurred for SWS when it is perceived as speech. This result indicates that the audiovisual integration process underlying the McGurk effect is speech-specific. To test the speech-specificity of the audiovisual detection advantage, we investigated if visual speech may assist auditory detection of SWS when perceived as non-speech and when perceived as speech (Eskelund *et al*., 2010).

METHODS

18 participants (six female), mean age 25 (range 21 to 30) all reported normal hearing and normal or corrected-to-normal vision. Four were excluded; three due to recognizing SWS as speech before entering the speech condition of the experiment and one due to not being able to discriminate among the SWS stimuli.

All stimuli were based on the speech recordings and SWS replicas produced by Tuomainen *et al*. (2005). Four auditory stimuli were used, SWS /omso/ and /onso/, and natural /omso/ and /onso/. A total of eight audiovisual stimuli were produced by combining SWS and natural speech tokens with video of the talking face, resulting in congruent and incongruent audiovisual combinations of /omso/ and /onso/.

In identification tasks, sound intensity of both SWS replicas was 77 dB SPL, while natural speech stimuli had intensities of 68 dB SPL and 70 dB SPL for natural /omso/ and /onso/ respectively. In detection tasks, a noise masker with a constant intensity of 65 dB SPL were added, while the intensity of the acoustic stimulus was varied, using a 2AFC paradigm and an adaptive staircase procedure. Duration of the masker was that of the stimulus plus two random intervals of 100-300 ms added before and after stimulus onset to eliminate any cues from onset of masker and target.

In the non-speech condition, subjects perceive SWS as non-speech sounds. Thus, when identifying and detecting audiovisual SWS tokens, they have little reason to look at the talking face, precluding any integration of sight and hearing. This might be a trivial confound for any reduction in the McGurk illusion in the non-speech condition. To control that subjects were actually looking at the screen, we included a secondary visual detection task. A white dot was overlaid the nose of the talking face for the same duration as each stimulus plus surrounding random intervals. In 20% of trials, the white dot disappeared for 200 ms at the onset of consonants /m/ and /n/. Subjects had to detect if the dot blinked.

Before the experiment, subjects were trained in discriminating the auditory SWS tokens /omso/ and /onso/ in arbitrary non-speech categories (‘sound 1’ and ‘sound 2’). The experiment began with a non-speech condition during which subjects were naïve about the speech origin of SWS. First, subjects identified SWS auditory-only, audiovisual congruent and audiovisual incongruent stimuli in arbitrary categories, then they performed the detection task with auditory-only and audiovisual congruent tokens of /omso/. After a short break, subjects were then informed about the speech-like nature of SWS and then followed the speech condition (Eskelund *et al*., 2010), repeating the identification and detection tasks, with the change in the identification task, that stimuli were now categorised as ‘omso’ and ‘onso’. Additionally, in the speech condition, a separate task of identifying natural auditory-only, audiovisual congruent and audiovisual incongruent speech tokens was performed.

As the experiment hinges on a shift in perceptual set between non-speech and speech perception, the hearing experience of participants was checked before and after each condition. Included subjects did not associate SWS with speech before being informed about its phonetic origin and they all reported hearing SWS as speech during all tasks in the speech condition.

RESULTS

Identification tasks

Proportions of correct responses in the identification of auditory stimuli are displayed in Figure 1 for each stimulus type. The results were subjected to an arcsine transformation and analyzed with a two-way (Stimulus x Conditions) repeated-measures ANOVA (Eskelund *et al.*, 2010). The interaction between Stimulus and Conditions was significant. This indicated that the effect of Condition differed for the three stimulus types.

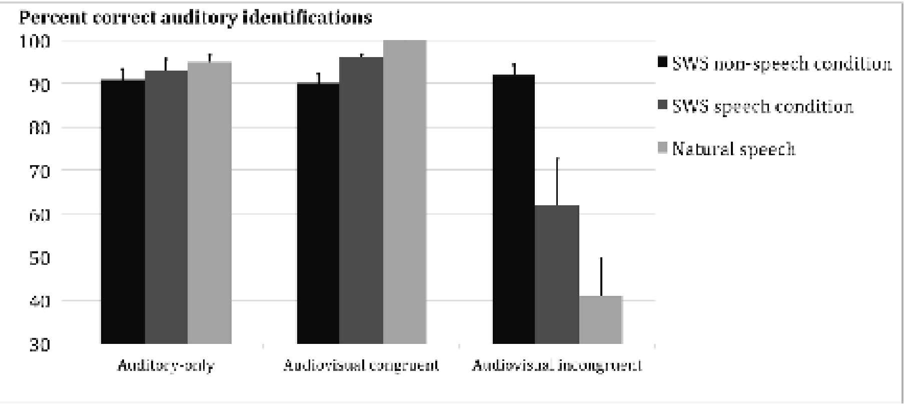

Fig. 1: Results from the auditory identification tasks. Bars represent percent correct auditory identifications by stimulus type and condition. Error bars represent standard error of mean. With audiovisual incongruent stimuli, the difference in identification performance between SWS in non-speech and speech conditions indicates that audiovisual integration of phonetic content only occurs in speech perception.

For auditory-only stimuli, there was no significant effect of Condition. In the case of congruent audiovisual speech, there was a significant effect of Condition. This indicated that performance was highest for natural speech, lower for SWS in the speech condition and lowest for SWS in non-speech condition. For congruent audiovisual stimuli, integrating the talking face with the voice should improve performance. Therefore this effect can be interpreted as a stronger influence from vision on audition when the stimulus is perceived as speech.

For incongruent audiovisual stimuli, which would tend to induce a McGurk illusion and hence is the pivotal stimulus class of the identification task, the effect of Condition was significant, reflecting that performance was lowest for natural speech, somewhat higher for SWS in the speech condition and highest for SWS in the non-speech condition. For audiovisual speech, seeing an incongruent talking face should obstruct identification due to the McGurk effect. This result could be interpreted as a stronger influence of vision on audition when SWS is perceived as speech.

A comparison of performance with auditory-only SWS, audiovisual congruent SWS and audiovisual incongruent SWS in the non-speech condition revealed no significant difference, indicating that the visual signal was not integrated into the auditory signal while perceived as non-speech.

Detection tasks

Detection thresholds were calculated as the mean of the last 10 responses of the adaptive staircase. Average thresholds are shown in Figure 2. Mean detection advantage of audiovisual stimuli over auditory-only presentation was 2.66 dB SPL. The detection difference between non-speech and speech conditions was negligible.

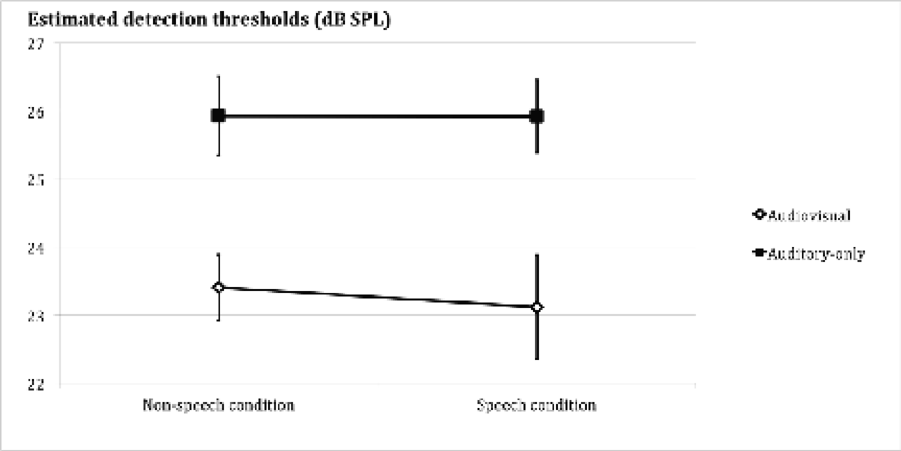

Fig. 2: Results from the auditory detection tasks. Points represent auditory detection threshold per stimulus type and condition. Error bars represent standard error of mean.

Results were subjected to a two-way (Stimulus x Conditions) repeated-measures ANOVA (Eskelund *et al.*, 2010). In contrast to the interaction seen in identification tasks, no significant interaction between factors Stimulus and Condition was found, indicating that the audiovisual detection advantage is not influenced by the shift from non-speech to speech perception. A significant main effect of Stimulus was found, however, expressing that a detection advantage for audiovisual SWS over auditory-only SWS occurred. No main effect of Condition was found.

Secondary task

Detection of the occurrence of the white dot remained consistently high across all tasks. No significant difference in secondary task performance was found between tasks (Eskelund *et al.*, 2010). This indicates that participants were following the instructions to look at the screen in all tasks, even in the non-speech condition where the talking face was irrelevant to tasks.

DISCUSSION

Identification results confirmed the observation of Tuomainen and colleagues (2005) that the McGurk illusion does occur for SWS but only when perceived as speech. This finding suggests that audiovisual integration of phonetic content is a speech-specific effect.

The audiovisual detection advantage observed for SWS in the present study is in concordance with Grant and Seitz' findings (2000) for natural speech. Interestingly, this effect was not influenced by whether SWS was perceived as non-speech or speech. The finding is in agreement with the interpretation of Bernstein and colleagues (2004), that the detection advantage is not specific for speech perception. In contrast, the McGurk effect only occurred in the speech condition. This suggests that the audiovisual detection advantage is not speech-specific whereas the McGurk effect is.

Our results thus further suggest that the detection advantage and the McGurk illusion are caused by two dissociated mechanisms, integrating different features of the audiovisual signal according to the perceptual set of the observer. This shows that audiovisual integration in speech perception is not, as Soto-Faraco and Alsius (2009) put it, a "monolithic", but rather a "multi-faceted" process.

Extending upon the concept of Auditory Scene Analysis (Bregman, 1990), Schwartz and colleagues (2004) proposed a two-stage model of audiovisual integration. In their concept of Audiovisual Scene Analysis, the early stage forms a correspondence between auditory and visual signals in a "primitive grouping" (Barker *et al*., 1998). This bimodal correspondence facilitates auditory detection by aiding segregation of auditory sources. Phonetic content is identified at a later stage, which receives the grouped bimodal signal.

In a recent series of experiments, Nahorna and colleagues (2011, 2010) showed that the illusory phonetic percept in the McGurk illusion could be disintegrated when the expectation of phonetic audiovisual congruence was changed. In one condition, subjects were presented with a series of congruent audiovisual syllables followed by an incongruent audiovisual syllable, which had to be identified. This produced a McGurk illusion. In a second condition, subjects were presented with a series of incongruent audiovisual speech syllables, again followed by an incongruent audiovisual syllable, which had to be identified. Now the McGurk illusion disappeared. According to the two-stage model, the early stage operates under the assumption of congruence, thus integrating unexpected incongruent auditory and visual signals as observed in the first condition. However, when evidence for audiovisual incongruence accumulates as in the second condition, the weight of the coherence evaluation in the early stage changes. As the grouping thus is reduced, the weight of the non-matching visual signal is decreased in the phonetic decision in the second stage. This results in the unbinding of the incongruent signals, eliminating the McGurk illusion. Our results agree with this approach. The early stage groups auditory and visual signals and facilitates auditory detection regardless of whether the listener is perceptually set for speech. In contrast, the integration of phonetic cues occurs in the later stage, which our results suggest is speech-specific.

Our current findings thus fit well with a multi-stage model as suggested by Schwartz (2004) and Nahorna (2011, 2010). An early stage would assess audiovisual coherence and exploit bimodal covariation to enhance the effective auditory signal-to-noise ratio. This is the stage involved in the audiovisual detection advantage. A later stage would identify phonetic content on basis of the percept generated in the first stage. This stage underlies the McGurk effect and is speech-specific.

REFERENCES

Barker, J. P., Berthommier, F., and Schwartz, J.-L. (**1998**). "Is Primitive AV Coherence an Aid to Segment the Scene?" In Proceedings of the International Conference on Auditory-Visual Speech Processing 1998 (Terrigal, Australia).

Bernstein, L., Auer, E. T. J., and Takayanagi, S. (**2004**)." Auditory speech detection in noise enhanced by lipreading". Speech Communication **44**, 5-18

Bregman, A. S. (**1990**). *Auditory scene analysis: the perceptual organization of sound* (MIT Press).

Eskelund, K., Tuomainen, J., and Andersen, T. S. (**2010**). "Multistage audiovisual integration of speech: dissociating identification and detection". Exp. Brain Res. **208**, 447–457.

Grant, K. W., and Seitz, P.-F. (**2000**). "The use of visible speech cues for improving auditory detection of spoken sentences". J. Acoust. Soc. Am. **108**, 1197-1208.

Massaro, D. W. (**1998**). *Perceiving talking faces: from speech perception to a behavioral principle* (MIT Press).

McGurk, H., and MacDonald, J. (**1976**). "Hearing lips and seeing voices". Nature **264**, 746–748.

Nahorna, O., Berthommier, F., and Schwartz, J.-L. (**2011**). "Binding and unbinding in audiovisual speech fusion: Follow-up experiments on a new paradigm". In Proceedings of the International Conference on Auditory-Visual Speech Processing 2011 (Volterra, Italy: Kungliga Tekniska Högskolan, Sweden).

Nahorna, O., Berthommier, F., and Schwartz, J.-L. (**2010**). "Binding and unbinding in audiovisual speech fusion: Removing the McGurk effect by an incoherent preceding audiovisual context". In Proceedings of the International Conference on Auditory-Visual Speech Processing 2010 (Hakone, Kanagawa, Japan: Kumamoto University, Japan).

Remez, R., Rubin, P., Pisoni, D., and Carrell, T. (**1981**). "Speech perception without traditional speech cues". Science **212**, 947–949.

Schwartz, J.-L., Berthommier, F., and Savariaux, C. (**2004**). "Seeing to hear better: evidence for early audio-visual interactions in speech identification". Cognition **93**, B69–B78.

Soto-Faraco, S., and Alsius, A. (**2009**). "Deconstructing the McGurk–MacDonald illusion". J. Exp. Psychol.: Human Perception and Performance **35**, 580–587.

Tuomainen, J., Andersen, T. S., Tiippana, K., and Sams, M. (**2005**). "Audio-visual speech perception is special". Cognition **96**, B13–B22.

Vatakis, A., and Spence, C. (**2007**). "Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli". Percept. & Psychophys. **69**, 744–756.

Objective measurement of listening effort while using first and second language in simulated cochlear implants

GANESH A C, VIJITHA SUNNY AND SUBBA RAO T A

Dr M V Shetty College of Speech and Hearing, Mangalore, India

It is generally believed that the cognitive effort to understand speech under adverse listening conditions differs between first and second languages. The present study examined this issue with 10 native Kannada speakers who use English as a second language. Subjects listened to noise band vocoder (simulated Cochlear Implant in normal's) processed sentences in quiet, noise (-6 dB SNR) and visual reaction time conditions. The listening effort was measured using a dual task paradigm. The mean scores obtained were better for Kannada than English. Repeated ANOVA measures indicated a significant main effect of listening conditions in both languages. The listening effort was larger while using English (second language). The visual reaction time data indicated a larger reaction time for English. The data in general suggests an increased cognitive effort for the processing of the second language. Speech perception under adverse listening conditions was significantly higher for the first language and demonstrates the importance of language proficiency in everyday listening conditions. The measurement of the listening effort using the dual-task paradigm has shown that it provides an additional index of speech perception under different listening conditions beyond traditional word recognition scores for each language in bilinguals.

INTRODUCTION

The communication process involves the transduction of acoustic signal to physiological information from the Auditory Periphery to the Cognitive System. It involves not only perceptual factors like the ability to hear, but also cognitive factors like listening, comprehending and responding (Kiessling *et al*. 2003). Hearing and listening are two different processes where most of the audiologists usually fail to clearly distinguish. Hearing is a sensory process and a passive function whereas; listening is an active process that demands attention and cognitive resources to understand speech.

The Listening effort is an essential dimension of speech understanding. It can be evaluated by using subjective and objective measurements in audiology clinical practice. Subjective measurements are self reports or rating scales e.g. Speech,

Proceedings of ISAAR 2011: Speech perception and auditory disorders. 3rd International Symposium on Auditory and Audiological Research. August 2011, Nyborg, Denmark. Edited by T. Dau, M. L. Jepsen, J. Cristensen-Dalsgaard, and T. Poulsen. ISBN 87-990013-3-0. EAN 9788799001330. The Danavox Jubilee Foundation, 2012.