

The effect of audiovisual speech training on the phonological skills of children with specific language impairment (SLI)

Child Language Teaching and Therapy
2018, Vol. 34(3) 269–287

© The Author(s) 2018

Article reuse guidelines:

sagepub.com/journals-permissions

DOI: 10.1177/0265659018793697

journals.sagepub.com/home/ctt



Jenni Heikkilä , **Eila Lonka**
University of Helsinki, Finland

Auli Meronen, Sisko Tuovinen, Raija Eronen
Valteri Centre for Learning and Consulting Services, Finland

**Paavo HT Leppänen, Ulla Richardson,
Timo Ahonen**
University of Jyväskylä, Finland

Kaisa Tiippana
University of Helsinki, Finland

Abstract

We developed a computerized audiovisual training programme for school-aged children with specific language impairment (SLI) to improve their phonological skills. The programme included various tasks requiring phonological decisions. Spoken words, pictures, letters and written syllables were used as training material. Spoken words were presented either as audiovisual speech (together with the talking face), or as auditory speech (voice alone). Two groups (10 children/group) trained for six weeks, five days per week: the audiovisual group trained with audiovisual speech, and the other group received analogically the same training but with auditory speech. Before and after training, language skills and other cognitive skills were assessed. The audiovisual group improved in a non-word-repetition test. Such improvement was not observed with auditory training. This result suggests that audiovisual speech may be helpful in the rehabilitation of children with SLI.

Corresponding author:

Jenni Heikkilä, Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, PO Box 9, Helsinki, 00014, Finland.

Email: jenni.heikkila@helsinki.fi

Keywords

audiovisual speech, rehabilitation, specific language impairment (SLI)

I Introduction

Visual articulatory gestures support speech perception in typically developed adults in all situations, but especially in noisy conditions (Ross, 2007; Sumby and Pollack, 1954; Zion Golumbic et al., 2013). Visual (V) gestures can even change the perception of auditory (A) speech signals, as demonstrated by the illusion known as the McGurk effect, in which dubbing an auditory consonant onto a video of incongruent articulatory movements alters auditory perception (McGurk and MacDonald, 1976). The change can be intermediate between the A and V consonants (A[b]V[g] heard as [d]), or it can correspond to the V consonant (A[b]V[d] heard as [d]) (McGurk and MacDonald, 1976; for a review, see Tiippana, 2014).

Audiovisual speech perception develops during childhood. Infants are able to match speech sounds and lip movements at just a few months old (Dodd, 1979; Kuhl and Meltzoff, 1982; Lewkowicz and Hansen, 2012). However, children from 6 to 10 years are still less sensitive to visual speech cues than adults (Massaro et al., 1986; McGurk and MacDonald, 1976; Ross et al., 2011; Sekiyama and Burnham, 2008).

Individuals with specific language impairment (SLI) experience problems in language skills. The term SLI refers to a developmental condition whereby a child does not develop his or her native language skills to the level of typically developing children at a typical pace, while the non-verbal and social skills are within a normal range (Law et al., 1998; Norbury et al., 2008; Tomblin et al., 1996, 1997). In SLI, it is usual for both expressive and receptive language skills to be delayed or to develop atypically. Problems with syntax, morphosyntax and verbal memory are often present, as well as problems with phonological processing (for a review, see Leonard, 2014). In addition, auditory perception and phonological memory and working memory are often impaired in SLI (for reviews, see Leonard, 2014; Webster and Shevell, 2004).

Children with SLI and more broadly defined developmental language disorders may also have difficulties with audiovisual and visual aspects of speech perception. Norrix et al. (2007) found that 4- to 5-year-old children with SLI have a weaker McGurk effect compared to typically developing children. Meronen et al. (2013) have shown that 8-year-old children with developmental language disorders have a weaker McGurk effect and also poorer performance in lip-reading. Furthermore, when the signal-to-noise ratio (SNR) of acoustic speech decreases, children with developmental language disorder do not rely as much on visual speech as is the case with typically developing children. The authors concluded that difficulties in using visual articulatory cues might contribute to less efficient speech perception, and that problems in the perception of visual speech may play a role in the development of language among children at risk of language impairment. Similarly, Leybaert et al. (2014) found that 8- to 14-year-old children with SLI identify audiovisually and visually presented consonants less accurately than their typically developing peers, and they attest to a weaker McGurk effect. Heikkilä et al. (2017) have shown that word-level lip-reading is impaired in 7- to 11-year-old children with SLI compared to typically developing children. They have further demonstrated that phonological skills are associated with lip-reading skills. In agreement, Knowland et al. (2016) have shown that both word-level and sentence-level speechreading is poorer in 5- to 11-year-old children with developmental language disorders compared to their typically developing peers. Taken together, previous studies show that children with SLI and other developmental language disorders are less influenced by visual aspects of speech than their typically developing peers.

Due to the general benefits of audiovisual speech for language comprehension, some previous studies have used audiovisual speech in training language skills. Shams et al. (2011) have argued that multisensory experiences can have an impact on subsequent unisensory processing. They have shown that motion perception improves more by audiovisual than visual training when motion direction is congruent in both modalities (Kim et al., 2008). There is neural evidence in support of this kind of a congruency benefit. For example, Li et al. (2011) showed that audiovisual presentation facilitates neural semantic access more than auditory presentation. In the same vein, the cortical representation of auditory speech is facilitated by congruent visual speech (Crosse et al., 2015), and temporal tracking of the speech envelope is enhanced by visual speech in noisy conditions (Zion Golumbic et al., 2013). Consequently, it has been hypothesized that audiovisual speech might be more efficient in training language skills than auditory speech alone. Bernstein et al. (2013) argued that visual speech provides guidance on learning phoneme distinctions because it is correlated with auditory features and contains associated phonemic cues. Moradi et al. (2017) have proposed that audiovisual training reinforces the routes to phonological and lexical representations, facilitating their later access even in unisensory conditions.

Several studies show that audiovisual speech can be helpful in training phonological skills and pronunciation in adults (Hardison, 2003; Hazan et al., 2005; Kawase et al., 2009; Lidestam et al., 2014; Richie and Kewley-Port, 2008). Phonetic contrasts (Hardison, 2003; Hazan et al., 2005) and pronunciation (Hazan et al., 2005) of a foreign language can be trained more effectively using audiovisual speech than auditory speech alone. For example, Hardison (2003) has shown that Japanese and Korean learners of English identify the /l/-/r/ phonemic contrast better after training with audiovisual speech than with auditory speech. Further, Hazan et al. (2005) have demonstrated that Japanese students perceive the /v/-/b/-/p/ phonemic contrast better with audiovisual speech than traditional auditory speech training alone. In addition, they have found that the pronunciation of consonants /l/-/r/ improved with audiovisual speech training. Similarly, Hirata and Kelly (2010) have shown that native English speakers learn Japanese phonetic contrasts (vowel durations) better with audiovisual than with auditory training. Training with audiovisual speech has also been beneficial in the rehabilitation of clinical populations. Adults with hearing impairment (Bernstein et al., 2014) and with aphasia (Fridriksson et al., 2009), as well as elderly hearing-aid users (Moradi et al., 2017) have been successfully trained with audiovisual speech. Fridriksson et al. (2009) trained speech production in patients with non-fluent aphasia by means of audiovisual speech. Patients used a computer-assisted picture-word matching task in which words were presented either as audiovisual or auditory speech. After training, picture naming improved in those patients who had received training with audiovisual speech compared to auditory speech. The results suggest that focusing on audiovisual speech can improve speech production, possibly due to a partly shared neural network recruited during audiovisual speech perception and speech production (Skipper et al., 2005). It should be noted, however, that not all studies with clinical populations show a training effect with audiovisual speech. Preminger and Ziegler (2008) trained participants with hearing loss to discriminate phonemes using audiovisual and auditory speech, but there was no improvement in speech perception measurements after training.

In the case of children, there are very few studies on training with audiovisual speech. Massaro and Light (2004) trained seven 8- to 13-year-old children with hearing loss using a computer-animated talking head. The talking head, called Baldi, was used as a language instructor for speech perception and pronunciation. Baldi spoke slowly and had transparent skin, allowing the articulatory movements inside the mouth to be seen. The children practiced several phonetic contrasts, including minimal pair words, and the production of speech segments. Exercises included distinguishing between voiceless and voiced consonants ('fan' vs. 'van'), and fricative and affricate sounds ('shoe' vs. 'chew'). The training sessions were carried out twice a week for 19 weeks, 15

minutes per lesson. All of the children improved in speech perception and production assessed by word identification and word repetition tasks (conducted by Baldi). The speech production training results were also generalized to words that had not been trained before. Bosseler and Massaro (2003) also studied Baldi's utility in teaching vocabulary and grammar to nine 7- to 12-year-old children with autism spectrum disorder (ASD). In ASD, the ability to produce and comprehend spoken language is often limited. With Baldi's help, the children were taught to associate pictures and spoken words. In the training tasks, Baldi orally instructed the children to select a picture presented on the computer screen (for example by saying 'click on the star'), and the children's task was to select an appropriate picture from among several pictures with a mouse click. Baldi also instructed the children to name the words. Visual feedback (pictures of happy or sad faces) was given for each response. The children trained with Baldi a few times a week over the course of six months. The training effect was studied immediately after the training and 30 days after the training by means of a word recognition test conducted by Baldi. The children learned a significant number of new words, and they were able to recognize most of the words 30 days after training. The children also generalized their vocabulary knowledge onto novel pictures that were not used during the training sessions, and transferred their vocabulary knowledge from the computer program to an independent assessment by an instructor. However, since there was no training with auditory speech only in the studies by Bosseler and Massaro (2003) and Massaro and Light (2004), it is impossible to confirm whether the training effect was due to audiovisual training.

More recently, Irwin et al. (2015) have continued studies in the utility of audiovisual speech on training spoken word perception for children with ASD. Children with ASD typically look less at speaking faces in the presence of auditory background noise than typically developing children (Irwin and Brancazio, 2014; Irwin et al., 2011). The hypothesis was that audiovisual training may help children with ASD to pay more attention to visual speech and to form stronger perceptual representations of spoken language. A computer application utilizing audiovisual speech was used for the training sessions. Four 8- to 10-year-old children with ASD participated in the training. The children were presented with video clips of a speaker uttering monosyllabic words with varying levels of auditory noise. After the video, four pictures were presented and each child selected the picture depicting the word meaning according to what s/he had heard. One picture matched the word, one distractor rhymed with the word, and others were selected to begin with consonants that differed in the place of articulation. In this study, the children played with the application for 3 days a week, for 10 minutes per session over a period of 12 weeks. Before and after training, they were tested with auditory words (50 words from the training set) presented at varying levels of noise. Accuracy in identifying the words increased after the training period. Training with audiovisual speech seemed to generalize to the auditory modality. The results, however, should be regarded as tentative because there was no training with auditory speech only, and just four children participated.

Audiovisual speech training has not been conducted previously in children with SLI. However, recently, Pedro et al. (2018) taught letter-sound correspondence to pre-schoolers with phonological delay using cards with visual cues (letters and speech sound production photographs) to stimulate letter-sound correspondences and speech sound production. In addition, Franceschini et al. (2017) found that playing action video games (commercial games that included tasks requiring temporal and spatial processing and motor planning) improved phonological decoding and speed of word recognition in children with dyslexia. These studies suggest that training with visual and video components might be beneficial for acquiring phonological skills in children with language learning difficulties.

The aim of the present study was to study the effect of audiovisual speech training on improving phonological skills of children with specific language impairment. Difficulties in language

acquisition can cause problems in school learning as well as communication in everyday life. Traditionally, children with SLI have been rehabilitated with speech therapy, yet computer-based training could enable easily available, cost-effective rehabilitation. The studies reviewed above suggest that audiovisual speech may be effective in training, at least in some clinical populations. We hypothesized that using audiovisual speech in phonological training tasks might improve the learning of phonological skills that are difficult for children with SLI. The audiovisual speech training programme contained phonological tasks and short-term memory tasks with varying levels of difficulty. Speech was presented both with and without auditory noise. Noise was added to encourage the children to rely more on visual speech cues and to train speech perception amid noise. For the training study, we used a programme utilizing audiovisual speech, and exactly the same programme utilizing auditory speech only, in order to particularly address the contribution of visual speech. We expected the programme involving audiovisual speech to be more effective in training phonological skills. Visual articulatory cues may help in the identification of phonemic structure, and thus children trained with audiovisual speech may develop more robust phonological representations compared to children trained with auditory speech only. Hence, we expected the children trained with audiovisual speech to improve more in phonological tasks than those trained with auditory speech. Children trained with audiovisual speech may also learn to utilize visual speech cues more effectively, which may be seen as an improvement in visual and audiovisual speech perception.

II Methods

I Participants

A total of 20 children, 7 girls and 13 boys (aged between 7.2 and 10.8 years, mean age 8.9 years), with an SLI diagnosis participated in the study. All were diagnosed by professionals (medical teams), and met the national diagnostic criteria for SLI, which are based on ICD-10 (Specific Language Impairment: Current Care Guidelines, 2010). According to the diagnostic criteria, the child's expressive and/or receptive language skills were markedly below (at least -2 SD) the average level for the child's age group, and at least -1 SD below the level for non-verbal skills. The non-verbal IQ was not below 70, as measured with standardized tests. For more detailed criteria, see Specific Language Impairment: Current Care Guidelines (2010).

Children with a comorbid neurological or psychiatric diagnosis (ADHD, autism spectrum disorders, general cognitive deficit, motor deficits) or hearing loss were excluded from the study. The children were recruited from Valteri Onerva School, which specializes in children who need special support due to difficulties related to vision, hearing or language. All participants had Finnish as their mother tongue, and their parents reported the children as having normal hearing, and normal or corrected to normal vision. The children had written permission from their parents to participate. The children were also informed that they could drop out of the study whenever they wanted to.

2 Stimuli

The audiovisual speech stimuli comprised video recordings of common Finnish nouns from several word categories. Six adult speakers were used: four females and two males. All were native speakers of Finnish and had clear articulation. In the video clips, speakers looked straight ahead and did not perform any movements other than speech. The faces were about 10 cm in height. In all of the clips, the word was pronounced slowly and clearly. Auditory noise was either pink (power

decreasing with increasing frequency) or babble noise (created by merging together five audio files of people talking). Noisy stimuli were created by merging noise into the audio files at a SNR of +5 dB for both noise types. In the training programme, the spoken words were presented without noise in 50% of the trials, with pink noise in 25%, and with babble noise in the remaining 25%. Noisy trials were presented randomly among clear trials to avoid the training sessions becoming too strenuous. The stimuli used for the response options comprised pictures, letters and syllables. The pictures were colorful drawings extracted from a picture database created and tested by Rossion and Pourtois (2004), or pictures drawn and modified to resemble those in the database. Some pictures were also taken from Pics for PECS images (www.pecs.com) with the permission of Pyramid Educational Consultants. Letters and syllables were written in 1.5 cm-high capital letters. The program was run with Presentation software (Neurobehavioral systems) using four laptop computers. The auditory stimuli were presented via headphones at 55 dB(A).

3 Training tasks

The programme contained four kinds of tasks: word-picture, word-letter, word-syllable and short-term memory tasks. The first three tasks were designed to practice phonological skills. All of the tasks were presented in blocks consisting of several trials presented in random order. In each task, an audiovisual video clip of a face uttering a word was presented. After the clip, four response options were shown: a match to the word, and three distractors. The child chose the option that matched what they had heard by clicking on it with a computer cursor received feedback for the response and proceeded to the next word. Feedback points were yellow smileys. Ten points were given after a correct answer. If the answer was incorrect, five points were given. The purpose of the feedback was to maintain motivation.

All of the tasks contained a large number of words with varying degrees of visual distinctiveness. In some cases, the visual articulation of the target word was very different from any of the distractor words, for example a bilabial target and non-labial distractors (such as target *pallo* = 'ball', distractors *kallo* = 'skull', *kello* = 'clock', *nukke* = 'doll'). In other cases, the visual articulations were not clearly distinctive (see, for example, Figure 1). We opted for this design to more closely simulate a natural situation, where visual distinctiveness varies freely. Moreover, it has been shown that visual speech utterances can be discriminated even when the articulations do not belong to separate viseme classes, for example when they do not differ in the place of articulation (Files et al., 2015).

a Word-picture tasks. In the word-picture tasks, an audiovisual clip was presented of a face uttering a word, and the task was to choose the picture that corresponded to the word. After the clip, four drawings were shown: one that matched the word, one that included a phonetic change to the word, and two other drawings (Figure 1). Different kinds of phonetic changes were used: a change in the first sound of the word, a change in the last sound of the word, a vowel change in the middle of the word, a consonant change in the middle of the word, a vowel duration change, a consonant duration change, a diphthong change, and a change in double consonant. There were 11 blocks of word-picture tasks, each containing 40 words.

b Word-letter tasks. In the word-letter tasks, an audiovisual clip was presented of a face uttering a word, and the task was to decide what the first letter of the word was. After the clip, four letters were shown: one that matched the first letter of the word, one that shared some phonetic similarity with the target letter (for example, if the target was 'M', the letter 'N' was presented), and two other letters. There were 10 blocks of word-letter tasks, each containing 40 words.



Figure 1. An example of the word-picture task and word-syllable task.

Notes. Word-picture task (above): Video still of the speaker uttering the word *takki* ('coat'), followed by four drawings: one that matches the word, one with a single phonetic change *nakki* ('sausage') and two other words *naula* ('nail') and *sukka* ('sock'). Word-syllable task (below): Video still of the speaker uttering the word *kukka* ('flower'), followed by four syllables: one that is the first syllable of the word (*kuk*), one that rhymes with the first syllable (*suk*) and two other syllables.

c Word-syllable tasks. In the word-syllable tasks, an audiovisual clip was presented of a face uttering a word, and the task was to decide what the first syllable of the word was (Figure 1). After the clip, four syllables were shown: one that matched the first syllable of the word, and three other syllables. One distractor syllable shared some phonetic similarity with the target syllable in order to make the task more challenging. There were six blocks of word-syllable tasks, each containing 40 words.

d Short-term memory tasks. In the short-term memory tasks, either two or three audiovisual clips of a face uttering a word were presented in sequence one after another, and the task was to memorize the words in the presented order. After the clips, four pictures were shown (2–3 correct and 2–1 distractors). The child clicked the pictures that corresponded to the words in the order presented. There were 10 blocks of short-term memory tasks, each consisting of 20 trials. Six blocks contained tasks with two words and four blocks contained tasks with three words. In the first two blocks, noisy trials were not presented so as to make the blocks easier.

e Evaluation methods. Before and after the training period, expressive and receptive language skills and other cognitive abilities were assessed with neuropsychological tests. Tests assessing phonological skills, verbal short-term memory, picture naming, attention and verbal motor skills were chosen because rehabilitation may have an effect on these. The training tasks were aimed at rehabilitating

phonological skills, but they may have also improved attentional skills because the tasks required sustained attention. Training with audiovisual speech might also influence speech production, and therefore affect verbal motor skills and picture naming (Fridriksson et al., 2009). Consonant discrimination was evaluated by measuring the percentages of correct responses to auditory and visual utterances [mi] and [ni], as well as their audiovisual combinations. The extent of visual influence on speech perception was assessed using the McGurk effect (McGurk and MacDonald, 1976). Lip-reading skills were assessed with a Finnish lip-reading test (Heikkilä et al., 2017).

f Neuropsychological tests. The participants' language skills, general cognitive abilities, attentional skills and working memory were assessed using standardized psychological tests. All tests apart from Raven Matrices and Imitating Hand Positions were also conducted after the training period.

Phonological skills were assessed using NEPSY-II Phonological Processing (Korkman et al., 2008) and NEPSY Repetition of Nonsense Words (Korkman et al., 1997). The Phonological Processing subtest evaluates the ability to perceive word structure, and consists of two tasks designed to assess phonemic awareness. Word Segment Recognition requires identification of words from word segments. Phonological Segmentation evaluates phonological processing at the level of word segments (syllables) and speech sounds (phonemes). The child is asked to repeat a word and then to create a new word by omitting a syllable or a speech sound, or by substituting one sound in a word for another.

Repetition of Nonsense Words measures the ability to analyse and reproduce phonological knowledge. In this subtest, the child is asked to repeat spoken nonsense words of varying length and complexity. The nonsense words were presented in the auditory modality using a tape recorder.

Verbal short-term memory was assessed with WISC-IV Digit Span (Wechsler, 2010) and NEPSY Repetition of Sentences (Korkman et al., 1997). In Digit span, a list of numbers is read aloud, and the task is to orally repeat the numbers. In Repetition of Sentences, the task is to repeat sentences of varying length and complexity.

Verbal comprehension and auditory short-term memory were assessed with NEPSY-II Comprehension of Instructions (Korkman et al., 2008), in which the child has to touch colored pictures according to increasingly complex oral instructions.

Picture-naming abilities and vocabulary were measured with the Boston Naming Test (Kaplan et al., 1983), in which the child is asked to name pictures. Verbal fluency was evaluated using NEPSY-II Word Generation (Korkman et al., 2008), in which the child is asked to say as many words as possible in a certain semantic (animals/foods) or phonetic category (the first letter is s/k) in one minute. Verbal motor skills were assessed with NEPSY Oromotor Sequences (Korkman et al., 1997), a subtest which measures verbal motor coordination and oral praxic functions, and in which the child is asked to repeat sequences of words and syllables. Manual sensorimotor skills were assessed with the NEPSY-II Imitating Hand Positions subtest (Korkman et al., 2008). In this subtest, the child imitates various hand positions presented by the examiner.

Attentional skills were evaluated with the NEPSY-II Visual Attention subtest (Korkman et al., 2008), which assesses the ability to focus and maintain attention on a visual target. In this task, the child searches for pictures of two target faces that are embedded among faces with varying features. Non-verbal intelligence was evaluated with Raven's Progressive Matrices (Raven et al., 1998).

g Consonant discrimination. Visual, auditory and audiovisual consonant discrimination was tested before and after training. Videos and/or audio files of a female speaker uttering meaningless syllables [ni] and [mi] were presented in random order. The task was to indicate whether the presented

consonant was [m] or [n] by pressing a button on a response device. In the visual task, videos were presented of the silently uttered syllable [ni] or [mi] (10 presentations of each). The duration of the visual syllables was 600 ms. In the auditory task, the auditory syllables [ni] and [mi] were presented (10 presentations of each). The duration of auditory syllables was 265 ms. During the presentation of the auditory syllables, a blurred picture was presented of the speaker's face. The audiovisual task included congruent audiovisual [mi] and [ni] syllables (10 of each) and 20 presentations of a McGurk stimulus which was created by dubbing an auditory syllable [mi] onto a video clip of the syllable [ni]. The McGurk effect occurs when the participant hears [ni] (e.g. MacDonald and McGurk, 1978). The same design and stimuli have been used previously with adults (Alsus et al., 2014; Tiippana et al., 2011). All of the children were assessed individually in a soundproof laboratory. Each child sat on a chair approximately 60 cm from a computer monitor where the visual stimuli appeared. Auditory stimuli were presented with loudspeakers at 75 dB(A). The tasks were presented in the same order, starting with a visual task, followed by an auditory task and an audiovisual task. Before each task, oral and brief written instructions were given to the children. The researcher sat next to the child during the task in order to ensure that he or she had understood the instructions and was oriented to the task, and to make sure that the child's gaze was directed towards the screen.

h Lip-reading test. Word-level lipreading skills were evaluated by means of a computer-based lipreading test (Heikkilä et al., 2017). The test included 17 Finnish words presented as silent video clips in which a female speaker uttered the words. After each video clip, four pictures were presented: one that matched the word, and three whose phonetic forms resembled those of the word seen in the video. The task was to lip-read the word and to select the corresponding picture by pointing it out to the researcher. Each child was assessed individually in a soundproof laboratory. The child sat on a chair approximately 50 cm from a laptop monitor where the stimuli appeared. The size of the face on the screen was 6.5 cm in height and 4 cm wide. The researcher sat next to the child in order to ensure that he or she had understood the instructions and was oriented to the task, and also to ensure that his or her gaze was directed towards the talking face.

4 Procedure

The participants were divided into two training groups, both of which comprised ten children: six boys and four girls (group mean age 8 years 9 months) in the audiovisual (AV) training group, and seven boys and three girls (group mean age 9 years 1 month) in the auditory (A) training group. Due to the small number of participants, the groups were matched according to age, gender and general cognitive skills rather than randomized. This was done in order to make the groups as similar and comparable as possible, which would have been quite impossible with random sampling. The audiovisual group received the training with the programme utilizing audiovisual speech. The auditory group received exactly the same training but with auditory speech only, so that when a spoken word was presented, a black computer screen was presented instead of the talking face. This design aimed to isolate the contribution of visual speech by comparing otherwise identical audiovisual and auditory training. The training period lasted six weeks. During that time, participants used the training programme on schooldays, 5 days a week, 10–15 minutes per day. Both training groups received the same amount of training. Short daily training sessions were conducted because it has been shown that 5- to 20-minute training sessions can be beneficial in the rehabilitation of language skills when repeated several times a week (Lovio et al., 2012). There was no time limit for responding, and the subsequent word was presented after the child had responded. One session consisted of one or two blocks, depending on the response speed of the participant. Two

speech therapists at the Valteri Onerva School supervised the training sessions and monitored participants to ensure that their gaze was directed towards the stimuli during training. The same speech therapists supervised both groups. The training session started with the word-picture tasks, followed by the word-letter tasks and the word-syllable tasks, concluding with the short-term memory tasks. The children completed all of the tasks except for three participants who had very poor letter knowledge who consequently did not do the word-letter and word-syllable tasks (two were in the audiovisual group and one in the auditory group). When it came to the memory tasks, the children started with the easier (two-word task) and progressed to the more challenging (three-word task). All of the children did the word-picture tasks twice. The children who did not do the word-letter and word-syllable tasks did the word-picture tasks three times.

Before and after the training period, expressive and receptive language skills and other cognitive abilities were assessed with neuropsychological tests and behavioral tasks. General cognitive abilities and motor skills were assessed before training. All assessments (both neuropsychological and behavioral) were conducted by the same researcher, who was a clinical neuropsychologist. At the time of testing, the training group membership was not known by the researcher. The assessments were conducted in the same order for all participants, starting with the neuropsychological assessment followed by the behavioral tasks. The subtests were presented in the same order in both pre- and post-testing. The neuropsychological assessment took about 60 minutes. The behavioral tasks took about 30 minutes. The neuropsychological and behavioral tests were conducted on different days. All testing was conducted during school days.

The training study was conducted in collaboration with the University of Helsinki, the University of Jyväskylä and Valteri Onerva School, Jyväskylä, Finland. During the training period, children with SLI used the programme at school during school days. The research has received ethical approval from the University of Helsinki Review Board in the Humanities and Social and Behavioral Sciences (statement 15/2013).

5 Statistical analyses

First, one way analysis of variance (ANOVA) with training group (AV, A) as a factor was used to ensure that there were no differences between the groups before training in age or performance in any neuropsychological tests and behavioral tasks.

Second, scores in the training programme were analysed by comparing different training tasks (word-picture, word-letter, word-syllable, short-term memory) and noise types (no noise, pink noise, babble noise) using t-tests in order to test whether the type of task or noise affected the scores during training.

The main interest in the analyses was to compare performance in different tests pre- and post-training. The training effect refers to an improvement in test performance after the training period. We were particularly interested in the audiovisual training effect. If there is an audiovisual training effect, the audiovisual training group should have significantly better test scores in post-measurement than in pre-measurement, while the auditory training group should not demonstrate this kind of improvement. The training effect was investigated by comparing pre- and post-measurements in neuropsychological tests and behavioral tasks between the audiovisual and auditory training groups. This was done by analysing performance in each test with a mixed model analysis of variance (ANOVA) with the testing time (pre-/post-training) and training group (AV, A) as factors. A main effect of testing time would indicate a training effect. An interaction of testing time and training group would indicate that the training effect differed between groups, confirming an audiovisual training effect if pairwise comparisons showed a significant improvement only in the AV group. For consonant discrimination, the

analysis additionally included a repeated-measures factor modality (A, V, AV) in order to test whether auditory, visual and congruent audiovisual consonants contributed to differences in recognition accuracy.

The Greenhouse–Geisser correction has been applied for p values when appropriate. However, the original degrees of freedom are reported, together with the epsilon values ϵ . Effect sizes are reported as partial eta squared. Bonferroni-corrected significance levels were used in pairwise comparisons with post-hoc t tests.

III Results

1 Comparison of groups before training

There were no differences between the audiovisual and auditory training groups in terms of age or any cognitive/behavioral test before training: Age ($F(1, 19) = .648, p = .430$), Raven Matrices ($F(1, 19) = .146, p = .707$), Repetition of Nonsense Words ($F(1, 19) = .48, p = .830$), Oromotor Sequences ($F(1, 19) = .38, p = .545$), Repetition of Sentences ($F(1, 19) = .226, p = .640$), Visual Attention ($F(1, 19) = .118, p = .292$), Comprehension of Instructions ($F(1, 19) = .331, p = .572$), Phonological Processing ($F(1, 19) = .520, p = .480$), Word Generation ($F(1, 19) = .64, p = .44$), Imitating Hand Position ($F(1, 19) = .550, p = .648$), Digit Span ($F(1, 19) = 1.3, p = .268$), Boston Naming Test ($F(1, 19) = .172, p = .683$), Lipreading ($F(1, 19) = 1.46, p = .243$), McGurk ($F(1, 19) = .76, p = .395$), Auditory syllables ($F(1, 19) = .466, p = .509$), Visual syllables ($F(1, 19) = 1.27, p = .274$), Audiovisual syllables ($F(1, 19) = .406, p = .502$).

2 Performance in training tasks

The percentages of correct responses in the training tasks for the audiovisual and auditory training groups are presented in Supplementary Material 1. Overall performance in the word-picture, word-letter and word-syllable tasks was good, with around 80–90% of trials correct. The performance between the audiovisual and auditory training groups did not differ significantly in any of the task types ($p > .05$ in every t -test comparison). Pink or babble noise did not degrade performance since no differences were observed between words without noise, words with pink noise and words with babble noise (averaged across groups) in word-picture and word-syllable tasks ($p > .05$ in every comparison). In word-letter tasks, however, performance was worse with pink noise compared to words without noise ($t(19) = 3.62, p = .002$). In short-term memory tasks, performance was poorer in three-word tasks than in two-word tasks ($t(18) = 5.99, p < .001$), but pink or babble noise did not disturb performance ($p > .05$ in every comparison).

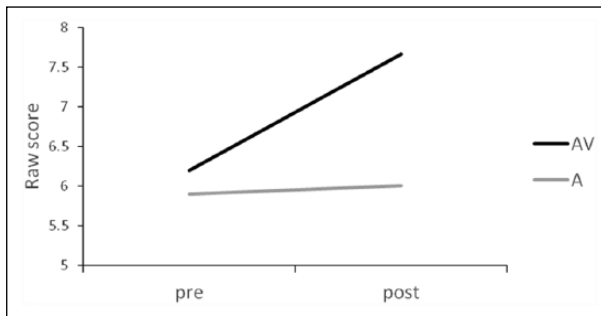
3 Training effects

a Standardized neuropsychological tests. The mean performances of the audiovisual and auditory training groups in the standardized tests before and after training are presented in Table 1 and Figures 2 and 3.

A statistically significant main effect of testing time ($F(1, 17) = 8.31, p = .002, \epsilon = 1.0$) and, importantly, an interaction of testing time \times training group ($F(1, 17) = 11.08, p = .004$, partial $\eta^2 = .395, \epsilon = 1.0$) was found only in the Repetition of Nonsense Words test. The interaction was due to an audiovisual training effect, in which the audiovisual training group improved significantly from pre- to post-measurement ($t(9) = 4.097, p = .006$), while the auditory training group did not ($t(9) = .361, p = .726$) (Figure 2).

Table 1. Mean raw scores and ranges in cognitive tests for both audiovisual (AV) and auditory (A) training group before (pre) and after (post) training.

	AV pre	AV post	A pre	A post
Boston Naming Test	36 (27–48)	39 (30–49)	37 (28–48)	40 (32–47)
Raven matrices	24.6 (13–32)		25.6 (14–33)	
Digit span	6.9 (3–13)	8.2 (5–12)	8.4 (3–12)	9.6 (7–12)
Repetition of nonsense words	6.2 (2–11)	7.66 (4–11)	5.9 (1–11)	6 (1–10)
Repetition of sentences	14.9 (8–19)	14 (7–20)	15.7 (10–21)	15.9 (8–22)
Oromotor sequences	32.4 (0–61)	40.3 (0–61)	37.3 (12–54)	42.4 (20–62)
Phonological processing	29.9 (14–48)	33.2 (24–45)	32.6 (25–44)	35.2 (29–45)
Comprehension of instructions	19.9 (16–23)	20.4 (18–25)	20.6 (18–26)	21.5 (17–25)
Word generation	24 (7–41)	27.3 (7–53)	28 (15–47)	26 (15–45)
Visual attention	2.6 (–11–+19)	7 (–13–+21)	7.1 (–9–+19)	9 (–2–+23)
Imitating hand positions	17 (12–21)		15–8 (7–20)	

**Figure 2.** Repetition of nonsense words.

Notes. Performance in the repetition of nonsense words test in the audiovisual (AV) and auditory (A) training group in pre-measurement and post-measurement. Audiovisual training improved performance ($p < .05$ between pre-and post-measurement in audiovisual group) while auditory training did not.

A main effect of testing time was found also in Phonological Processing ($F(1, 17) = 10.58$, $p = .005$, partial $\eta^2 = .384$, $\epsilon = 1.0$), Boston Naming Test ($F(1, 17) = 5.15$, $p = .037$, partial $\eta^2 = .232$, $\epsilon = 1.0$), Digit Span ($F(1, 17) = 4.47$, $p = .049$, partial $\eta^2 = .208$, $\epsilon = 1.0$), and Oromotor Sequences ($F(1, 17) = 10.16$, $p = .005$, partial $\eta^2 = .374$, $\epsilon = 1.0$). In these tests, performance improved in both groups after training.

No main effect of testing time was found in Repetition of Sentences ($F(1, 17) = .024$, $p = .878$, partial $\eta^2 = .001$, $\epsilon = 1.0$), Visual Attention ($F(1, 17) = 3.15$, $p = .095$, partial $\eta^2 = .164$, $\epsilon = 1.0$), Comprehension of Instructions ($F(1, 17) = .568$, $p = .461$, partial $\eta^2 = .032$, $\epsilon = 1.0$) and Word Generation ($F(1, 17) = .171$, $p = .685$, partial $\eta^2 = .010$, $\epsilon = 1.0$) (Figure 3). No testing time \times training group interaction was found in any test other than Repetition of Nonsense Words; not significant tests: Digit Span ($F(1, 17) = .049$, $p = .757$, partial $\eta^2 = .006$, $\epsilon = 1.0$), Oromotor Sequences ($F(1, 17) = .001$, $p = .975$, partial $\eta^2 = .00$, $\epsilon = 1.0$), Boston Naming Test ($F(1, 17) = .007$, $p = .936$, partial $\eta^2 = .000$, $\epsilon = 1.0$), Phonological Processing ($F(1, 17) = .012$, $p = .916$, partial $\eta^2 = .001$, $\epsilon = 1.0$), Repetition of Sentences ($F(1, 17) = .169$, $p = .686$, partial $\eta^2 = .010$, $\epsilon = 1.0$), Visual Attention ($F(1, 17) = .087$, $p = .386$, partial $\eta^2 = .052$, $\epsilon = 1.0$), Comprehension of

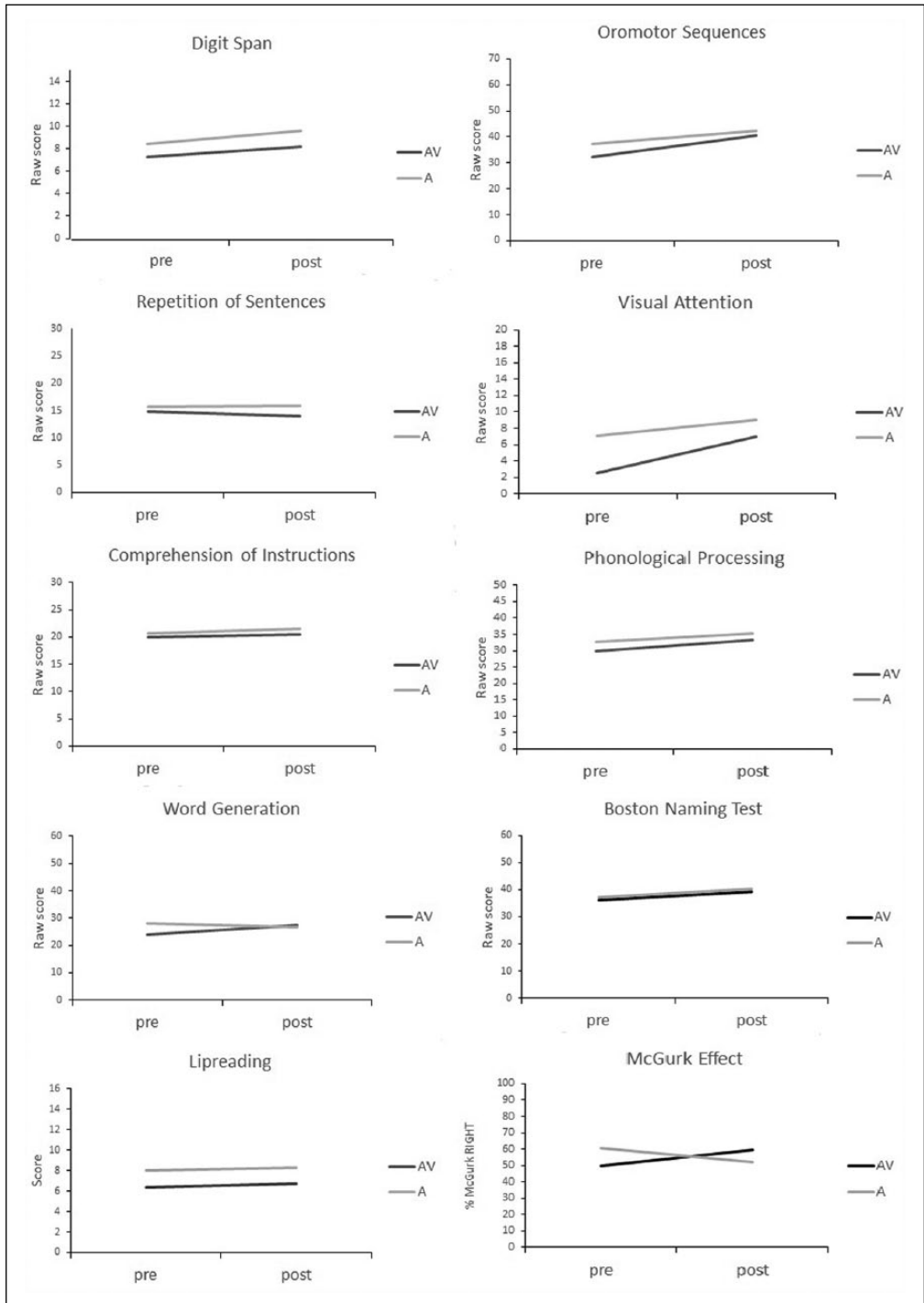


Figure 3. Performance in neuropsychological tests and behavioral tasks in the audiovisual (AV) and auditory (A) training group in pre-measurement and post-measurement.

Table 2. Mean performances and standard deviations in consonant discrimination test (percentages of correct responses) and in lipreading test (lipreading scores) for both audiovisual (AV) and auditory (A) training group before (pre) and after (post) training.

	AV pre	AV post	A pre	A post
Audiovisual congruent consonants	75.5% (17.7)	70% (25.7)	80.5% (28.3)	91.5% (7)
Audiovisual incongruent consonants (McGurk) according to vision	50.0% (30.1)	40.5% (35.5)	40.0% (23.2)	48.0% (34.8)
Visual consonants	68.0% (23)	66.0% (17.3)	64.5% (22.9)	77.5% (17.4)
Auditory consonants	75.5% (17.4)	71% (17.6)	78.5% (24)	82% (24.5)
Lipreading test	6.4 (3.5)	6.7% (2.8)	8 (2.2)	8.3 (1.8)

Instructions ($F(1, 17) = .207, p = .655, \text{partial } \eta^2 = .012, \epsilon = 1.0$), Word Generation ($F(1, 17) = 1.00, p = .331, \text{partial } \eta^2 = .056, \epsilon = 1.0$).

b Consonant discrimination. The consonant discrimination performance scores before and after training are presented in Table 2. Only modality had a statistically significant main effect ($F(2,18) = 5.17, p = .020, \text{partial } \eta^2 = .223, \epsilon = .737$). The syllables were more difficult to recognize in the visual modality compared to the audiovisual modality ($t = (19) = -2.8, p = 0.035$) but not compared to auditory modality ($t = (19) = -1.99, p = 0.18$). No differences were observed between audiovisual and auditory modality ($t = (19) = -1.2, p = 0.72$). No main effect of testing time was observed ($F(2,18) = .480, p = .497, \text{partial } \eta^2 = .026, \epsilon = 1$) and there was no interaction, which means that neither group improved during training ($F(2,18) = 2.85, p = .108, \text{partial } \eta^2 = .137, \epsilon = 1$). For the McGurk effect, there was no main effect of testing time ($F(1,18) = 1.9, p = .185, \text{partial } \eta^2 = .096, \epsilon = 1.0$), nor an interaction ($F(1,18) = .006, p = .940, \text{partial } \eta^2 = .001, \epsilon = 1.0$) because the strength of the McGurk effect did not change in either training group during training (Figure 3).

c Lip-reading test. Performance in the lip-reading test before and after training is presented in Table 2 and Figure 3. There was no main effect of testing time ($F(1,18) = .367, p = .552, \text{partial } \eta^2 = .020, \epsilon = 1.0$), nor was there an interaction between testing time and training group ($F(1,18) = .00, p = 1.0, \text{partial } \eta^2 = .00, \epsilon = 1.0$) showing that lip-reading did not improve during the training period in either group.

IV Discussion

Our results suggest that training with audiovisual speech can improve the phonological skills of children with SLI, at least in the repetition of nonsense words. The children who used the training programme utilizing audiovisual speech improved in the Repetition of Nonsense Words, whereas children in the auditory training group did not. This is the first study showing that audiovisual speech might be more effective than auditory speech in training phonological skills in children. Previous studies have shown training effects with audiovisual speech, but a comparison with an auditory training programme has been lacking (Bosseler and Massaro, 2003; Irwin et al., 2015; Massaro and Light, 2004).

Performance improved in both training groups in Phonological Processing, the Boston Naming Test, Digit Span and Oromotor Sequences. This suggests that both audiovisual and auditory training enhanced phonological skills and other verbal skills as well as working memory. Training also

included recognition of pictures, which might have facilitated picture naming in the Boston Naming Test. Since there was no control group who received no training in this study, these improvements might also be seen as reflecting a rehearsal effect. No general rehearsal effect can be attested, however, since not all tests exhibited a benefit at re-test.

The current findings imply that adding visual speech cues to phonological training might improve in particular those cognitive skills that are related to a nonword repetition task. Previous studies show that nonword repetition tasks require several cognitive skills: speech perception, phonological encoding, phonological memory, phonological assembly and articulation (for a review, see Coady and Evans, 2006). Children with SLI usually have deficits in several of these skills, leading to poor performance in nonword repetition. It has been suggested that because of systematic below-average performance in this task, nonword repetition could be a reliable marker for differentiating between children with SLI and children with typical language development (Coady and Evans, 2006). The process of accurately repeating a nonword involves perceiving the word, creating at least a transient phonological representation in the working memory, segmenting the new string into speech units, marking the temporal order of the units, formulating a motor plan for articulation and then implementing the plan (Snowling et al., 1991). If some of these processes fail, the repetition of a nonword is inaccurate. Training with audiovisual speech might improve several skills required in successful nonword repetition. Seeing the speaker's mouth provides meaningful visual information that is related to the accompanying sounds. Visual speech provides information about speech segments that are both redundant and complementary to the auditory signal. This can make the speech sounds more salient for processing (Bernstein et al., 2013; Hirata and Kelly, 2010).

We suggest that seeing the speaker's articulatory movements during training might help children to identify ambiguous speech sounds and thus enhance their phonological representations. Phonological representation is a term that describes the storage of phonological information in the long-term memory. Phonological skills such as phonological awareness and the ability to analyse and reproduce phonological knowledge (as in a nonword repetition task) reflect phonological representations (Elbro and Jensen, 2005; Fowler, 1991; Swan and Goswami, 1997). Less stable phonological representations may cause difficulties in assembling both the phonological and the articulatory codes for nonwords. Seeing the articulatory gestures during the training tasks may help in differentiating the speech sounds and in forming more robust phonological representations that may help in analysing and segmenting the speech units of spoken words. Phonological training using audiovisual speech might therefore strengthen the phonological representations more than auditory training, as reflected in better performance in the Repetition of Nonsense Words task.

In addition to strengthening phonological representations, audiovisual speech training might also help to formulate and implement more precise articulatory plans. Training unfamiliar phonetic contrasts with audiovisual speech improves not only the perception but also the production of these contrasts (Hazan et al., 2005). Speech perception, either by listening to auditory speech or by looking at articulatory movements, enhances the excitability of the motor areas related to speech production (Watkins et al., 2003). Furthermore, audiovisual speech activates cortical motor areas related to planning and executing speech production more than auditory speech alone (Skipper et al., 2005). We suggest that because of the connection between audiovisual speech perception and speech production, audiovisual training might also improve articulatory planning. Children with SLI may have difficulties in articulatory planning when trying to create an accurate version of a nonword that they need to repeat. Audiovisual speech training might reinforce the children's phonological representations and skills to form an accurate articulatory plan, to which end the improvement in both of these skills can be seen in the Repetition of Nonsense Words test. The test requires both of these skills, while the other tests related to articulation and phonological awareness

(Oromotor Sequences, Phonological Processing) do not necessarily do so. This might explain why the improvement was observed only in the Repetition of Nonsense Words test.

In addition, nonword repetition tasks have also been suggested to measure phonological short-term memory (Newbury et al., 2005). Our training programme also contained short-term memory tasks, which may have partially contributed to the improvement in the Repetition of Nonsense Words task. In adults, working memory performance improves when digits are presented using audiovisual speech compared to auditory or visual speech (Frtusova et al., 2013). Previous studies with children show that recognition memory (Heikkilä and Tiippana, 2016) and the learning of word lists (Murray and Thomson, 2011) improve in school-aged children with typical language development when stimuli to-be-memorized are presented audiovisually. These studies, however, did not include memory training, and hence it is not known whether memory skills can be trained using audiovisual methods.

No improvement was found in the consonant discrimination tasks. For congruent audiovisual consonants, this may simply be due to a ceiling effect, as performance was already high before training. This finding is not very surprising for lip-reading either, since the auditory words could always be heard rather well, to the extent that pure lip-reading was never required during training. On the other hand, if audiovisual training enhances the use of visual speech cues, it could be expected that the McGurk effect might be strengthened. Nonetheless, this did not happen. The original study by McGurk and MacDonald (1976) has shown that the effect is stronger in adults than children, suggesting that long-term experience increases visual influence. Unfortunately, there are no previous studies which would show whether any kind of training can influence the McGurk effect. Perhaps a much longer training period than the one used in the current study would be needed.

There are limitations in this study. This study was conducted with a rather small number of participants, which limits the strength of conclusions that can be drawn. Also, the lack of a control group without any training is a limitation. Further studies with a larger number of children with SLI and with a control group with no training would be extremely important.

V Conclusions

The results of this study suggest that training with audiovisual speech can improve the phonological skills of children with SLI more than training with auditory speech, at least in the case of the repetition of nonsense words. Audiovisual speech training might help children to create more robust phonological representations and improve their ability to plan and execute articulation.

Acknowledgements

We are grateful to the Valteri Centre for Learning and Consulting Services, Jyväskylä, Finland, and to the pupils and teachers in Valteri Onerva School, Jyväskylä, Finland, where the training study was conducted. We thank Radostina Georgieva for her help in video editing, Sanna Ahola for her help in creating and piloting the training programme, and Otto Loberg for his help with the behavioral data acquisition.

Declaration of conflicting interest


The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by grants from the Arvo and Lea Ylppö Foundation, Avohoidon

tutkimussäätiö Foundation, Emil Aaltonen Foundation, Finnish Brain Foundation, Finnish Concordia Fund, Otto Malm Foundation and University of Helsinki.

ORCID iD

Jenni Heikkilä  <https://orcid.org/0000-0002-2531-1273>

References

- Alsius A, Möttönen R, Sams ME, Soto-Faraco S, and Tiippana K (2014) Effect of attentional load on audio-visual speech perception: Evidence from ERPs. *Frontiers in Psychology* 5.
- Bernstein LE, Auer Jr. ET, Eberhardt SP, and Jiang J (2013) Auditory perceptual learning for speech perception can be enhanced by audiovisual training. *Frontiers in Neuroscience* 7: 34.
- Bernstein L, Eberhardt S, and Auer E (2014) Audiovisual spoken word training can promote or impede auditory-only perceptual learning: Prelingually deafened adults with late-acquired cochlear implants versus normal hearing adults. *Frontiers in Psychology* 5: 934.
- Bosseler A and Massaro D (2003) Development and evaluation of a computer-animated tutor for vocabulary and language learning in children with autism. *Journal of Autism and Developmental Disorders* 33: 653–71.
- Coady J and Evans J (2008) Uses and interpretations of non-word repetition tasks in children with and without specific language impairments (SLI). *International Journal of Language and Communication Disorders* 43: 1–40.
- Crosse MJ, Butler JS, and Lalor EC (2015) Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. *J Neuroscience* 35: 14195–204.
- Dodd B (1979) Lip reading in infants: attention to speech presented in-and out-of-synchrony. *Cognitive Psychology* 11: 487–84.
- Elbro C and Jensen M (2005) Quality of phonological representations, verbal learning and phoneme awareness in dyslexic and normal readers. *Scandinavian Journal of Psychology* 46: 375–84.
- Files BT, Tjan BS, Jiang JT, and Bernstein LE (2015) Visual speech discrimination and identification of natural and synthetic consonant stimuli. *Frontiers in Psychology* 6: 878.
- Fowler A (1991) How early phonological development might set the stage for phoneme awareness. *Haskins Laboratories Status Report on Speech Research* SR 105/106: 53–64.
- Franceschini S, Trevisan P, Ronconi L, et al. (2017) Action video games improve reading abilities and visual-to auditory attentional shifting in English-speaking children with dyslexia. *Scientific Reports* 7: 5863.
- Fridriksson J, Baker J, Whiteside J, et al. (2009) Treating visual speech perception to improve speech production in nonfluent aphasia. *Stroke* 40: 853–58.
- Frtusova J, Phillips N, and Winneke A (2013) ERP evidence that auditory-visual speech facilitates working memory in younger and older adults. *Psychology and Aging* 28: 481.
- Hardison D (2003) Acquisition of second-language speech: effects of visual cues, context and talker variability. *Applied Psycholinguistics* 24: 495–522.
- Hazan V, Sennema A, Iba M, and Faulkner A (2005) Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication* 47: 360–78.
- Heikkilä J, Lonka E, Ahola S, Meronen A, and Tiippana K (2017) Lipreading ability and its cognitive correlates in typically developing children and children with specific language impairment (SLI). *Journal of Speech, Language and Hearing Research* 60: 485–93.
- Heikkilä J and Tiippana K (2016) School-aged children can benefit from audiovisual semantic congruency during memory encoding. *Experimental Brain Research* 234: 1199–1207.
- Hirata Y and Kelly SD (2010) Effects of lips and hands on auditory learning of second-language speech sounds. *Journal of Speech, Language, and Hearing Research* 53: 298–310.
- Irwin JR and Brancazio L (2014) Seeing to hear? Patterns of gaze to speaking faces in children with autism spectrum disorders. *Frontiers in Psychology* 5: 397.
- Irwin JR, Preston J, Brancazio L, D'Angelo M, and Turcios J (2015) Development of an audiovisual speech perception app for children with autism spectrum disorders. *Clinical Linguistics and Phonetics* 29: 76–83.

- Irwin J, Tornatore L, Brancazio L, and Whalen D (2011) Can children with autism spectrum disorders 'hear' a speaking face? *Child Development* 82: 1397–1403.
- Kaplan E, Goodglass H, and Weintraub S (1983) *Boston Naming Test*. Philadelphia, PA: Lea and Febiger.
- Kawase T, Sakamoto S, Hori Y, et al. (2009) Bimodal audio-visual training enhances auditory adaptation process. *Neuroreport* 20: 1231–34.
- Kim RS, Seitz AR, and Shams L (2008) Benefits of stimulus congruency for multisensory facilitation of visual learning. *PLoS ONE* 3: e1532.
- Knowland V, Evans S, Snell C, and Rosen S (2016) Visual speech perception in children with language learning impairments. *Journal of Speech, Language and Hearing Research* 59: 1–14.
- Korkman M, Kirk U, and Kemp SL (1997) *NEPSY, Lasten Neuropsykologinen Tutkimus [The assessment of neuropsychological abilities in children]*. Helsinki: Psykologien Kustannus.
- Korkman M, Kirk U, and Kemp SL (2008) *NEPSY-II, Lasten Neuropsykologinen Tutkimus [The assessment of neuropsychological abilities in children]*. Helsinki: Psykologien Kustannus.
- Kuhl P and Meltzoff A (1982) The bimodal perception of speech in infancy. *Science* 218: 1138–41.
- Law J, Boyle J, Harris F, Harknes A, and Nye C (1998) Screening for speech and language delay: A systematic review of the literature. *Health Technology Assessment*, 2: 1–184.
- Leonard L (2014) *Children with specific language impairment*. Cambridge, MA: MIT Press.
- Lewkowicz DJ and Hansen-Tift A (2012) Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences* 109: 1431–36.
- Leybaert J, Macchi L, Huyse A, et al. (2014) Atypical audio-visual speech perception and McGurk effects in children with specific language impairment. *Frontiers in Psychology* 5.
- Li Y, Wang G, Long J, et al. (2011) Reproducibility and discriminability of brain patterns of semantic categories enhanced by congruent audiovisual stimuli. *PLoS ONE* 6: e20801.
- Lidestam B, Moradi S, Pettersson R, and Ricklefs T (2014) Audiovisual training is better than auditory-only training for auditory-only speech-in-noise identification. *Journal of the Acoustical Society of America* 136: EL142–47.
- Lovio R, Halttunen A, Lyytinen H, Näätänen R, and Kujala T (2012) Reading skill and neural processing accuracy improvement after a 3-hour intervention in preschoolers with difficulties in reading-related skills. *Brain Research* 1448: 42–55.
- MacDonald J and McGurk H (1978) Visual influences on speech perception processes. *Perception and Psychophysics* 24: 253–57.
- Massaro DW and Light J (2004) Using visible speech to train perception and production of speech for individuals with hearing loss. *Journal of Speech, Language, and Hearing Research* 47: 304–20.
- Massaro DW, Thompson LA, Barron B, and Laren E (1986) Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology* 41: 93–113.
- McGurk H and MacDonald J (1976) Hearing lips and seeing voices. *Nature* 264: 746–48.
- Meronen A, Tiippana K, Westerholm J, and Ahonen T (2013) Audiovisual speech perception in children with developmental language disorder in degraded listening conditions. *Journal of Speech, Language and Hearing Research* 56: 211–21.
- Moradi S, Wahlin A, Hällgren M, Rönnberg J, and Lidestam B (2017) The efficacy of short-term gated audio-visual speech training for improving auditory sentence identification in noise in elderly hearing aid users. *Frontiers in Psychology* 8: 368.
- Murray J and Thomson M (2011) Age-related differences in cognitive overload in an audio-visual memory task. *European Journal of Psychology of Education* 26: 129–41.
- Newbury D, Bishop D, and Monaco A (2005) Genetic influences on language impairment and phonological short-term memory. *Trends in Cognitive Science* 9: 528–34.
- Norbury CF, Thomplin JB, and Bishop D (2008) *Understanding developmental language disorders: From theory to practice*. Oxford: Psychology Press.
- Norrix L, Plante E, Vance R, and Boliek C (2007) Auditory-visual integration of speech in children with and without specific language impairment. *Journal of Speech, Language and Hearing Research* 50: 1639–51.

- Pedro C, Lousada M, Hall A, and Jesus L (2018) Visual stimuli in intervention approaches for pre-schoolers diagnosed with phonological delay. *Logopedics Phoniatrics Vocology* 43: 20–31.
- Preminger J and Ziegler C (2008) Can auditory and visual speech perception be trained within a group setting? *American Journal of Audiology* 17: 80–97.
- Raven J, Raven JC, and Court JH (1998) *Coloured progressive matrices*. San Antonio, TX: Pearson.
- Richie C and Kewley-Port D (2008) The effects of auditory-visual vowel identification training on speech recognition under difficult listening conditions. *Journal of Speech, Language, and Hearing Research* 51: 1607–19.
- Ross LA, Molholm S, Blanco D, et al. (2011) The development of multisensory speech perception continues into the late childhood years. *European Journal of Neuroscience* 33: 2329–37.
- Ross L, Saint-Amour D, Leavitt V, Javitt D, and Foxe J (2007) Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex* 17: 1147–53.
- Rossion B and Pourtois G (2004) Revisiting Snodgrass and Vanderwart’s object set: The role of surface detail in basic-level object recognition. *Perception* 33: 217–36.
- Sekiyama K and Burnham D (2008) Impact of language development of auditory-visual speech perception. *Developmental Science* 11: 306–20.
- Shams L, Wozny DR, Kim R, and Seitz AR (2011) Influences of multisensory experience on subsequent unisensory processing. *Frontiers in Psychology* 2: 264.
- Skipper J, Nusbaum H, and Smalle S (2005) Listening to talking faces: motor cortical activation during speech perception. *Neuroimage* 25: 76–89.
- Snowling M, Chiat S, and Hulme S (1991) Words, non-words and phonological processes: Some comments on Gathercole, Willis, Emslie and Baddeley. *Applied Psycholinguistics* 21: 369–73.
- Specific Language Impairment. Current Care Guidelines (Käypä Hoito) (2010) Working group set up by the Finnish Medical Society Duodecim, Finnish Phoniatrians (Suomen Foniatri Ry) and Finnish Association of Pediatric Neurology. Helsinki: The Finnish Medical Society Duodecim. Available at: www.kaypahoito.fi (accessed July 2018).
- Sumby WH and Pollack I (1954) Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America* 26: 212–15.
- Swan D and Goswami U (1997) Phonological awareness deficits in developmental dyslexia and the phonological representation hypothesis. *Journal of Experimental Child Psychology* 66: 18–41.
- Tiippana K (2014) What is the McGurk effect? *Frontiers in Psychology* 5: 725.
- Tiippana K, Puharinen H, Möttönen R, and Sams M (2011) Sound location influences audiovisual speech perception when spatial attention is manipulated. *Seeing and Perceiving* 24: 67–90.
- Tomblin J, Records N, and Buckwater P (1997) Prevalence of specific language impairment in kindergarten children. *Journal of Speech, Language and Hearing Research* 40: 1245–60.
- Tomblin J, Records N, and Zhang X (1996) A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech, Language and Hearing Research* 39: 1284–94.
- Watkins K, Strafella A, and Paus T (2003) Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia* 41: 989–94.
- Webster R and Shevell M (2004) Neurobiology of specific language impairment. *Journal of Child Neurology* 19: 471–81.
- Wechsler D (2010) WISC-IV: Wechsler Intelligence Scale for Children. Finnish edition. Helsinki: Psykologien Kustannus.
- Zion Golumbic E, Cogan GB, Schroeder CE, and Poeppel D (2013) Visual input enhances selective speech envelope tracking in an auditory cortex at a ‘Cocktail Party’. *Journal of Neuroscience* 33: 1417–26.