László Hunyadi, István Szekrényes, László Czap & István Sziklai: Seeing the sounds?
Argumentum 10 (2014), 325-338
Debreceni Egyetemi Kiadó

László Hunyadi, István Szekrényes, László Czap & István Sziklai

Seeing the sounds?[*]

Abstract

The paper presents the first results of a series of experiments aimed at assisting deaf people in improving their speech production and perception. The theoretical novelty of the approach is that it is based on the expected effect of brain plasticity. It is assumed that speech can be transcoded into visual patterns in such a way that the resulting stream of patterns will both help the learner in acquiring proper pronunciation and in processing the transcoded sounds online as continuous speech. The results shed light on some of the similarities and differences between auditory and visual perception. Further research is needed to possibly reconcile the apparent conflict between richness of visual information for better speech pronunciation and reduction of visual redundancy for better visual sound perception.

Keywords: deafness, auditory and visual perception, audiovisual transcoding, speech processing

1 Introduction

In recent years, in public discourse, through the initiatives of individuals as well as national and international programs we have been experiencing a growing need for raising the awareness of the majority society to offer equal chances to minorities, especially those with some kind of disabilities or vulnerabilities. This is due to actual inability to control congenital disabilities in spite of vast developments in genetics and molecular biology during the past decade. Indeed, coming from the majority we may often forget that many of our everyday physical abilities and capabilities may not simply be taken for granted, something as given, but such that being deprived of them can lead to social disabilities of the individual and the disfunction of society with regard to establishing a caregiving network in general. It is especially obvious that lacking the primary senses, vision or hearing, can directly affect the integration of such people in the society often resulting in lower level of education and lower economic and social status. A milestone in the habilitation and rehabilitation of deafness is cochlear implantation. The vast majority of cases with congenital and acquired deafness belong to the category of those hearing losses that can be restored by cochlear implantation (Zahnert 2011). Conversely, a significant part of the world's deaf population resists hearing habilitation for cultural, social, social security and healthcare reasons. These deaf individuals prefer, instead, sign language (Weisleder 2012, Baertschi 2013). Sign language, however, in

[*] This research was partly supported by Project No. TÁMOP 4.2.2-C/11/1/KONV-2012-0002.

the community of hearing individuals is inappropriate for everyday communication in the era of informatics. Whereas it is important to raise the awareness of each and every individual to such problems, it is especially important to augment and combine the efforts of those specialists who can contribute to solutions diminishing some of the social and individual effects of these disabilities. The authors of this paper, physicians, linguists and engineers have the aim to bring the world of sounds closer to the deaf so that they can pronounce speech sounds and speak in a natural way and, ultimately, can perceive speech sounds and synthesize them into the perception of connected speech.

It is widely known that the deaf undergo education in designated schools whose primary aim is to teach them to communicate with the majority society through acquiring the sounds, words and grammar of the target language of the society. This education has a very long history: back in time the ancient Egyptians interacted with them through hieroglyphs and gesturing, the two major methods used until the present. The natural language of the deaf, accordingly, is sign language. In today's practice the deaf learn to read and write and through the writing system (and assisted by studying the visual patterns of articulation) they acquire the sounds, the minimal building blocks of speech. This is the bilingual education of the deaf (signing and speaking). To acquire proper pronunciation is, however, far from being easy and the resulting sounds are usually restricted approximations of the spoken sounds due to the fact that finer details of their pronunciation are hidden from visible articulation. The mapping of the writing system onto sounds is also problematic: usually based on a combination of phonetic, phonological and historical principles it does not offer a single direct clue to pronunciation (not to mention pictographic or ideographic writing systems which are even more difficult in this respect). A hybrid writing system with its need for multiple coding/decoding is not an efficiently suitable basis for teaching speech sounds. In addition, even if one could pronounce a speech sound close to its original pattern, it does not, by itself, guarantee the understanding of the word it is part of. What is still missing is the sound > phoneme mapping, i.e. the generalization of a set of different sounds as representations of a single phoneme. It is a complex cognitive task forming the basis of speech perception; this is the process that allows us to only perceive those differences in sounds that are significant (meaningful) in comprehension (van Muenster & Baker 2014). This is why we have a challenge of two extremes: the teaching of pronunciation should be realized as close to the target as possible down to the minute details, whereas those details which are not significant (not meaningful) should be disregarded.

Still, in order for the deaf to have the chance to properly produce target speech sounds they need some sample that they will attempt to link to pronunciation. Following the insufficiency and inefficiency of representation by either the hybrid writing system or the existing visual articulation we came to the following suggestion: using an audiovisual transcoder we should give speech sounds a visual representation in the form of videograms and, through training, use these videograms as visual stimuli to produce the sensation of speech sounds. In order to facilitate proper, close to natural pronunciation, these videograms need to be linked in our cognitive system to the corresponding sounds (Hickok et al. 2011): we designed a transparent talking head that, eliminating the disadvantages of lip reading, demonstrates the articulation of the given speech sound in all its necessary details. Accordingly, based on the effect of a putative cross modal sensory motor integration (Bavelier & Neville 2002) we expect that the necessary link between the visually exposed sound and its motor articulation will be established, as a result of which the sensation of the sound through vision will take place (Finney et al. 2001, Kang et al. 2006, Merabet & Pascual-Leone 2010).

We expect that the rich visual representation of sounds through audiovisual transcoding will enable the deaf learner to observe the sound in minute details and, with online feedback of his/her own pronunciation the learner will adjust the articulation accordingly, ultimately enhancing the production of speech sounds.

As mentioned above, proper pronunciation (sounds) cannot effectively be separated from perception (interpretation as phonemes). Accordingly, we wish to find out to what extent audiovisual transcoding is suitable for the online perception and subsequent adequate synthesis of sequences of sounds, i.e. words or even longer speech segments. As a matter of fact, when teaching individual sounds or even words, each as a single videogram, to the deaf, the task is to observe the pattern in the videogram and adjust one's own articulation according to the visual feedback from the videogram and the transparent talking head. Whereas this use of the audiovisual transcoding is based on pattern recognition and respective adjustments to it, its use for perception includes an additional factor, that of processing time. We can assume that there is a direct relation between richness of information and processing time: namely, whereas teaching pronunciation requires information-rich videograms in that, in turn, require longer processing time, it may well turn out to be the case that this processing time is too long for online perception in which case different transcoding schemes might be necessary for the two processes of production and perception. (In actual fact, there is an apparent conflict between the need for rich visual information for better speech pronunciation and a corresponding need for redundancy free visual information for better visual processing. For redundancy in vision, cf. I. Kovács 2010.)

To our best knowledge ours is the first proposal for an audiovisual transcoder to represent speech sounds via visual perception, and no previous studies have compared auditory and visual perception in the function of speech motor processing (Hickok & Poeppel 2007). The primary basic question we wanted to answer is whether visual speech motor integration does exist. Therefore, we carried out a series of tests to find out how the generated videograms can be representative of speech sounds and how much processing time their identification requires. We also wanted to find out whether the visual processing of sound patterns is sequential in nature and whether we can synthesize (generate) new sequences from already learned ones, as potential properties of both audio and visual perception. The tests were carried out with hearing subjects, and the first results will be incorporated in the administering of teaching tests with deaf children in the near future.

Before we describe and analyze the tests with hearing subjects, the next two sections will present the specifications of the audiovisual transcoder and the experimental settings, respectively.

2 The theoretical basis for and practical implementation of the audiovisual transcoding and videograms

The theoretical basis of the audiovisual transcoder for the representation of sounds as a visual signal for the deaf is supported by studies showing that the integration in the brain of both acoustic and visual patterns co-emerging during the production of speech sounds significantly increases intelligibility. Similarly, this integration significantly enhances automatic speech recognition as well (Potamianos, Neti & Deligne 2003). In the case of people with hearing loss, the stronger the acoustic signal distortion, the more they rely on the visual modality of speech. We expect to be able to use vision as an additional component of this integration in two ways: first, subjects (practically people with hearing loss) can use as a visual stimulus the visual representation of articulation (something like reading the lips, but going further, seeing even the hidden areas of articulation employing a transparent talking head), and second, by extending this visual stimulus to a visual representation of the sound wave itself. The idea of our audio-visual transcoding is then to encode the main features of speech sounds into abstract visual signals – which we call videograms – and represent them graphically.

We have defined two types of videograms: a matrix construct representing momentary sound properties, and a column construct synchronously visualizing longer sound sequences: sound combinations (syllables), words and sentences.

As for the matrix construct, the actual sound wave of the speaker is represented by a predefined number of squares, each of them assigned to discrete frequency components. According to the actual frequency value, variation of these squares by colour conveys information about wavelength: bass sounds (having longer wavelength) are painted by colours having corresponding longer wavelength (red), treble sounds having shorter wavelength are painted by colours of shorter wavelength (blue, magenta). As a basis for transcoding the audio frequency range of 125 Hz–8000 Hz was selected, the one with primary significant for speech perception. This range is divided into a predefined number of frequency bands according to the so-called Bark scale (Zwicker & Terhardt 1980). Each of the bands is associated with a corresponding sector of the display, and the transcoded sounds in visual form are plotted onto the appropriate positions in these sectors, with their sizes and colours corresponding to the transcoded acoustic information. Being based on measured frequency and intensity values, the audovideo transcoder visualizes the individual sound differences within and across speakers.

See Diagram 1 below showing the matrix representation of the sound /i/ (window size is 40 ms – i.e. the minimum duration within which sound frequencies and intensities are processed and subsequently represented as a single frame of the videogram):

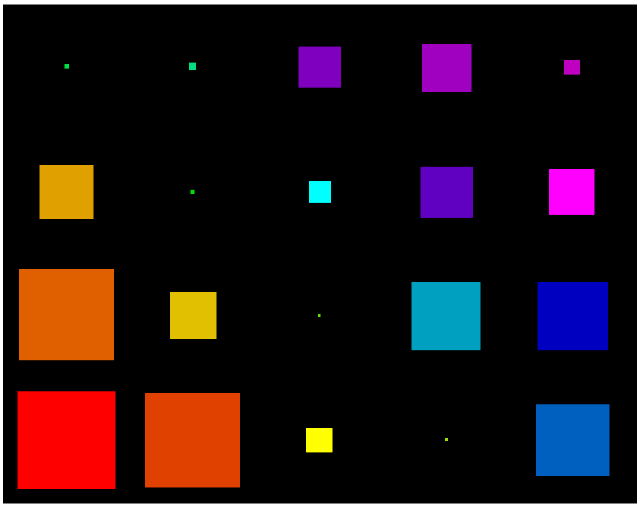

Diagram 1: Matrix representation of the sound [i]

As for the column representation, basically the same transcoding is applied to a sequence of frames so that, in contrast to the matrix construct, several frames can be simultaneously seen: with each new sample the previous frame moves further across the screen and remains visible (remember, the matrix construct differs in that every frame representation is replaced by the next frame representation). Accordingly, we expect that the column construct with more actual presentation time for each of the frames may prove to be more suitable for longer sequences, including words or sequences of words. In addition to colour and position, we introduced a third code here: a rectangle (the basic unit of column representation) also varies in size for intensity.

For this purpose, we have transformed matrices to columns, enabling the representation of a number of frames. During speech training lessons, comparing the current utterance of the trainee to the pre-recorded reference one is being demanded. A neural network based voice activity detector is trained with 4.5 hours of voices of more than 300 speakers. Acoustic samples for voice activity detection have been recorded by using different microphones, sound cards and PC-s, in a variety of noisy office, laboratory and home environment. After the current sample has been recorded, the beginning and end of speech is detected. Its transcoded column representation is displayed in a window above that of the reference one for us to be able to compare them. For the column representation of voice we introduced a third code in addition to position and colour. In addition to varying by size, a rectangle also varies its colour intensity in order to highlight the important segments. Diagram 2 below shows the resulting column videogram:

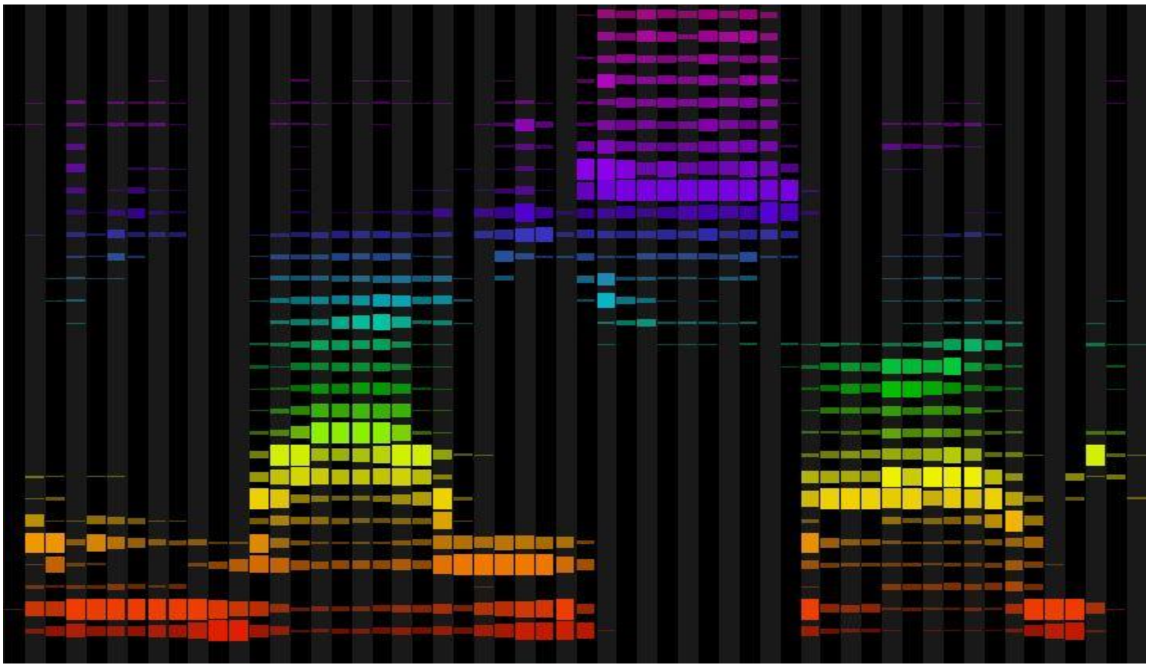

Diagram 2: Column representation of the word riválissal [rivAliS:Ol] 'with a rival'

As a first step in sound processing, sounds were recorded at a sample frequency of 16 kHz. After digitization the Fourier transform was applied, and further filtering according to the Bark scale was used to transform the given sound into a frequency specific visual pattern. The transcoder processes the sound in real time making it possible to process larger segments of speech in real time. As an effect, real time transcoding provides the learner with the possibility to compare in real time the videogram of the master sound sample with that of the subject's pronunciation. The learning process is further enhanced by the supplementary application of the transparent talking head, which makes it possible for the subject to adjust their lip and tongue movements to achieve a close approximation to the sample pronunciation.

3 Experiments with hearing subjects: An online interface for learning the videograms

The first round of teaching and testing involved the most common three vowels and four additional consonants (see Section 4). The sounds and their consonant + vowel sequences were recorded by 19 students (10 male and 9 female voices, aged 20-25 years, with no apparent speech disorders). For administering the tasks a web based client-server interface was designed with an installable client on the learners' own computer. The participants in the first rounds of the teaching included 19 university students from Debrecen and 10 from Miskolc. Since we wished to test which of two different videogram constructs may be more suitable for learning the sounds through audio video transcoding, half of the students were assigned one construct (matrix) and half another construct (column), and the groups changed after two weeks to study the construct previously studied by the other half. Each of the constructs was represented by two variants that only differed in orientation.

As for the process of teaching, each student listened to the recordings of each of the sounds/sound combinations by 19 different speakers accompanied by the respective videogram. Each set of listening was initiated by pressing a key on the keyboard. In the next round students could repeatedly listen to the sound and view the videogram by clicking on the visual construct. The interface allowed for moving to the test phase after 1000 clicks. For testing the same interface was used and the timing also remained the same: five rounds of testing in five days. The test consisted of the same selected 60 patterns for each learner. Learners were not given any prior information about the underlying principle of encoding, accordingly they had to find their own rules to identify each of the different sounds/sound combinations and disregard individual speaker variations of the same sound/sound combination.

4 Test results and analysis

4.1 Test 1: Single vowels and consonant + vowel sequences

The participants of the test were 18 hearing university students aged 20-25 years. The learning and testing environment was web based and the process followed a predefined order with a predefined number of trials within a predefined period. The learning phase was followed by online testing: subjects were supposed to type the sound representation of each of the ideograms presented.

The first test was aimed at identifying the videogram (sub)type with the highest number of correct responses. The patterns consisted of videograms of the following 9 sounds (3 vowels and 6 consonants), the basic vowels and a selection of consonants which might appear difficult for recognition and, consequently, require special attention for transcoding: [a:], [u], [i], [f], [v], [ʃ], [ʒ], [s], [z].

The selection of these sounds also followed general properties of human language development (MacWhinney 2005) and Pléh (2006) as well as pedagogical considerations (Bánréti 1979, Gósy 2005). For the articulation of each of the sounds in the test we consulted Molnár (1973).

The types of videograms were matrix and column ones, the subtypes only differing in the direction of the layout (horizontal vs. vertical). As Figure 1 shows, the individual videograms slightly differed in the number of presentations.

This first test shows that videograms with the column format proved to be somewhat more successfully recognised, but the success rate was quite low in all four types (cf. Figure 1 and Figure 2) (legend: x axis = transcoding types where subtypes A and B only differed in horizontal/vertical orientation of the pattern; y axis: number/% or responses, as shown)

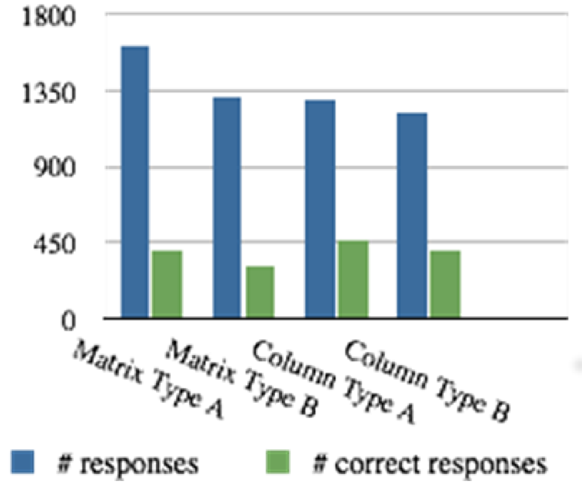

Figure 1: Test 1

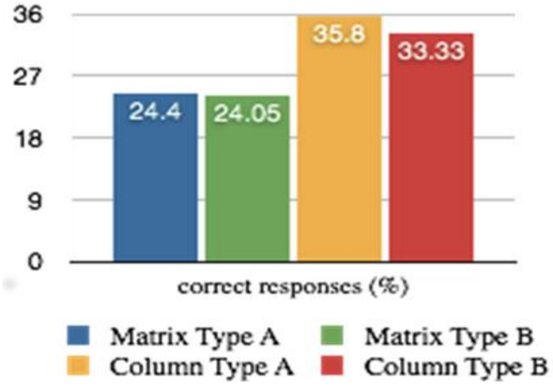

Figure 2: Test 1. Videogram type and correct responses

We wished to find out if this difference was also reflected in their respective recognition. As a comparison of Figures 3 and 4 below demonstrate, the successful recognition of vowels was much higher than that of consonants:

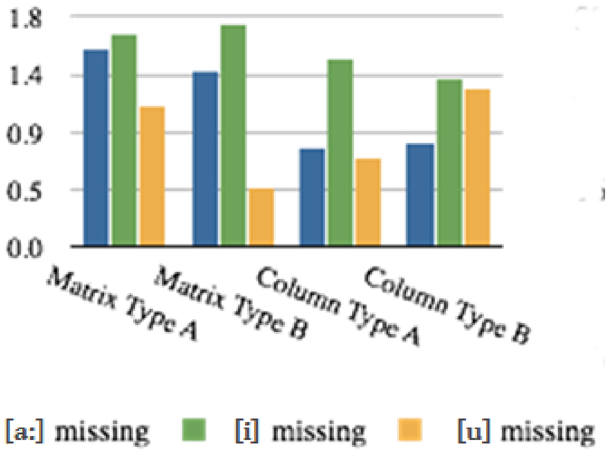

Figure 3: The pattern starts with a vowel. The recognition of the vowel fails (%)

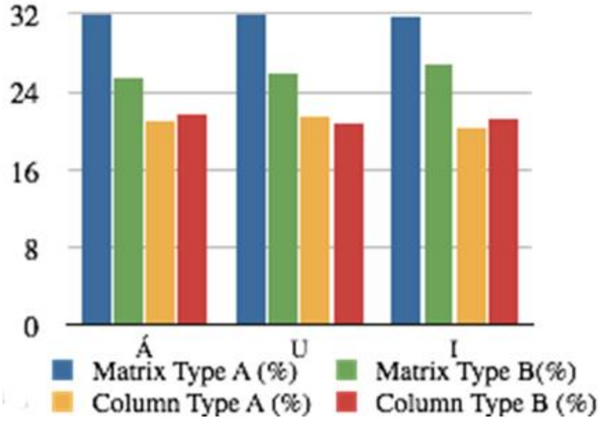

Figure 4: The pattern starts with a consonant. The recognition of the consonant fails before vowels (%)

We can notice in the above comparison that the matrix types perform better in the case of initial vowels, whereas the column types prove to be better in the case of initial consonants.

Compare now Figures 5 and 6 to find out how a misrepresentation of the pattern initial sound (vowel or consonant) bears an effect on the recognition of the following sound (consonant and vowel, respectively):

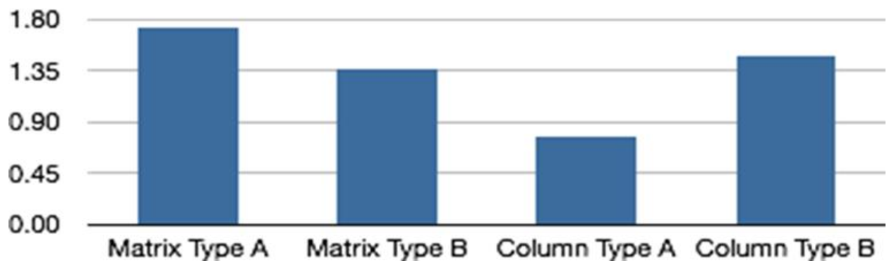

Figure 5: The pattern starts with a vowel. The response starts with a consonant and the following vowel is also incorrect (%)

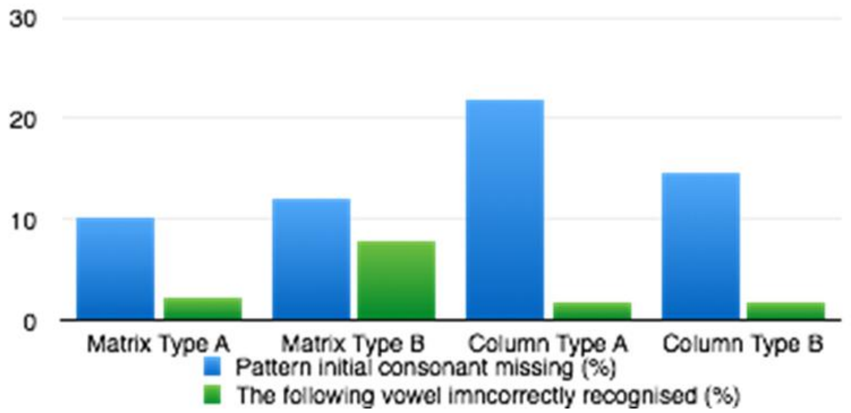

Figure 6: The pattern initial consonant missing and the following vowel incorrectly recognised (%)

Figure 5 above confirms yet again that the recognition of vowels is more successful than that of consonants: according to Figure 5, it is only in very rare cases that the initial vowel is not perceived at all (the recognized pattern starts with a consonant) and that the vowel following this incorrect consonant is also different from the missed initial vowel.

Figure 6 above demonstrates again the better recognition rate of vowels in another context. Even in cases when 10-20 % of the pattern initial consonants is missing in the recognition, the percentage of vowels incorrectly recognized is much lower. Comparing the performance of the matrix and column types, we can observe that there is again an asymmetric relation between the two: when the pattern initial sequence of consonant + vowel is concerned, the matrix performs better in the case of the initial consonant, but the column is found better in the case of the following vowel.

Although this first series of tests was carried out under conditions less stringent than the ones described in section 3 above (each subject used his/her individual web environment for learning the patterns and testing the correctness of responses without feedback in the learning phase and without self-evaluation), for initial orientation it is interesting to note reaction times (the time elapsed from pattern presentation to responding by typing in the answer) for each of the four rounds of tests (two different pattern orientations both for the matrix and for the column construct), shown in Table 1 (total number of trials: 1618 (Matrix Type A and Matrix Type B, 1619 (Column Type A and Column Type B)):

Videogram	Mean	Std Dev	Sum	Minimum	Maximum
Matrix Type A	5.2962	10.1923	8569.44	0.439	276.133
Matrix Type B	4.0746	8.4843	6591.01	0.57	242.21
Column Type A	4.4039	3.601	7129.86	0.712	39.304
Column Type B	3.5083	3.3006	5679.96	0.643	65.293

Table 1: Reaction times for tests involving matrix and column representations

This table demonstrates the processing time of the subjects in learning the definitely very unusual visual patterns and their corresponding sound equivalents (pairwise correlation was significant at $p < 0.03$ or better): the minimum value around or just above half a second may eventually suggest an objective lower limit to recognition, whereas both the high maximum values and the corresponding standard deviation may be the reflection of individual differences due to several possible factors (differences in study time, motivation, learning strategies etc.), factors that we wished to better control in the tests to follow. The table shows that the matrix constructs required comparably longer processing time than the column ones.

Since we could not rely on any previous experience about the speaker independence of the videograms generated, we also wished to find out to what extent recordings of the same sound patterns and their translation into videograms could actually represent sounds abstracted from individual speaker differences, an issue essential to sound and speech perception in general. The subjects learned the videograms as generated from recordings of 19 different speakers. The actual tests only included 6 speakers' recordings equally distributed along two groups of patterns as follows (cf. Table 2):

speaker group	**speaker ID**	**pattern**
1	1, 2, 18	[a:], [u], [i], [f], [v], [ʃ], [ʒ], [s], [z]
2	3, 8, 14	[fa:], [fi], [fu], [va:], [vi], [vu], [ʃa:], [ʃi], [ʃu],], [ʒa:], [ʒi], [ʒu], [sa:], [si], [su], [za:], [zi], [zu]

Table 2: Distribution of speakers across groups of patterns

4.2 Test 2: Testing the sequentiality of vision using pseudowords

The idea behind this test was the following: if subjects are able to learn the videograms to this relatively high degree of correct recognition, let us find out if they can identify a sequence of two previously learned patterns presented to them as a single (complex) pattern even if they had not previously seen them. We assumed that if this kind of synthesis was successful, it could serve as an argument for segmentation in visual perception (similarly to the temporal segmentation in audio perception). The pseudowords included the following (all composed of sounds and syllables from the first experiment): [fa:fu], [fufa:], [fa:vu], [vufa:], [va:fu], [fuva:], [ʒiʃi], [ʃisi], [susu], [fiu] both in matrix and column format.

What we observed first during the test was that the subjects felt totally lost seeing the videograms. It suggested to us that their learning process was highly demanding in general, requiring many repetitions of the same pattern rather than some involuntary generalisations of the mind. The results show that our impressions were correct; cf. Tables 3 and 4:

Subjects	**Correct: all sounds**	**Correct: 1st sound**	**Correct: 2nd sound**	**Correct: 3rd sound**	**Correct: fourth sound**	**Correct: 1st + 2nd sound**	**Correct hits in a single pattern (all patterns)**
JI_m	3	7	8	4	7	6	26
KD_m	0	6	7	1	5	5	19
KS_m	0	8	8	0	3	7	19
NA_m	1	7	7	1	3	6	18
NCSZS_m	1	4	6	2	5	4	17

VK2_m	0	4	6	2	5	2	17
LZS4_m	0	3	8	0	5	2	16
LL_m	0	2	5	4	4	1	15
SZS_m	0	7	7	0	1	6	15
BI_m	0	6	3	2	2	3	13
SA_m	0	3	3	3	4	1	13
SJ_m	0	2	4	3	4	1	13
VEGM_m	1	4	2	3	3	1	12
MegyN3_m	0	5	3	0	2	2	10
SZG_m	0	3	3	0	4	1	10
GA_m	1	2	2	1	1	2	6
VM_m	0	3	1	0	1	1	5
MCS_m	0	2	1	0	0	0	3

Table 3: Pseudowords. Correct recognition across subjects (matrix)

Subjects	**Correct: all sounds**	**Correct: 1st sound**	**Correct: 2nd sound**	**Correct: 3rd sound**	**Correct: forth sound**	**Correct: 1st + 2nd sound**	**Correct hits in a single pattern (all patterns)**
JI_c	3	7	8	5	8	5	28
KD_c	4	8	6	6	6	6	26
KS_c	3	6	8	4	5	5	23
MegyN3_c	3	7	7	5	4	5	23
LZS4_c	1	5	7	4	6	3	22
NCSZS_c	2	4	8	4	6	3	22
NA_c	0	4	8	2	3	4	17
VK_c	1	2	9	2	3	2	16
SZS_c	0	7	7	1	0	6	15
SA_c	0	6	4	2	2	3	14
VEGM_c	0	3	4	4	2	1	13
BI_c	0	4	4	2	2	2	12
SJ_c	0	2	4	2	3	0	11
SZG_c	0	0	5	2	4	0	11
VM_c	0	4	3	1	2	2	10

LL_c	0	4	1	3	0	0	8
MCS_c	0	4	1	1	1	1	7
GA_c	0	1	0	0	0	0	1

Table 4: Pseudowords. Correct hits across subjects (column)

Since the test included 10 patterns each for matrix and column, the maximum number of hits for each position in the pseudoword was 10. We can see from the results that in most of the cases the correct hits fell on the first half (the first two sounds). Also, in most of the cases the fourth sound received more correct hits than the preceding, third sound. It suggests an interesting strategy for perception: one tries to synthesize the sound patterns even if they did not come across with the exact pattern before. However, probably due to memory restrictions, the degree of recognition for the later sounds is smaller. Interestingly in this respect, however, the last sound may receive more “attention” than the preceding one. Even though our data may not be sufficient for a broader generalization, this phenomenon may prove to be similar to the recognition (reading) of written words: in the latter, too, the first and last letters are more prone to perception than the ones in between. Here, however, we cannot base our explanation on our familiarity with the visual form of the pattern as a whole, since the given pattern was totally unfamiliar to the subjects. The reason for this interesting behaviour of the perception of our videograms may still require further considerations.

5 Conclusions

(1) Visual integration of speech perception and speech motor exists in young normal hearing listeners as a possible cross-modal alternative of natural auditory integrated speech perception and production

(2) Exhibition of longer duration segments (250 msec vs 40 msec) of the speech transcoded videograms is beneficial for recognition by normal hearing individuals

(3) The orientation of videograms does not significantly influence pattern recognition

(4) The success of videogram recognition differs as a function of the selected transcoding strategy: better recognition by matrix videograms for syllables beginning with a vowel versus better recognition by spectrographic videograms for syllables beginning with a consonant

(5) Further investigation needed whether actual performance by young normal hearing individuals can (at least partly) be influenced by visual memory (mid term, 2-3 months in our tests) regarding speech videograms

(6) Further research into the nature of audio and visual perception is needed to possibly reconcile the apparent conflict between richness of visual information for better speech pronunciation and reduction of visual redundancy for better visual sound perception.

References

Baetschi, B. (2013): Hearing the implant debate: therapy or cultural alienation? *Journal International de Bioethique* [International journal of bioethics] 24(4), 71-81.

Bánréti, Z. (1979): *Gyerek és anyanyelv* [Child and Mother Tongue]. Budapest: Tankönyvkiadó.

Bavelier, D. & Neville, H.J. (2002): Cross-modal plasticity: Where and how? *Nature Reviews Neuroscience* 3, 443-452.

Finney, E.M., Fine, I. & Dobkins, K.R. (2001): Visual stimuli activate auditory cortex in the deaf. *Nature Reviews Neuroscience* 4(12), 1171-1173.

Gósy, M. (2005): *A beszédészlelés és a beszédmegértés fejlesztése iskolásoknak – Szülők számára* [The Deveplopment of Student's Speech Perception and Understanding – for Parents]. Budapest: Nikol, 2005.

Hickok, G. & Poeppel, D. (2007): The cortical organization of speech processing. *Nature Reviews Neuroscience* 8, 393-402.

Hickok, G., Houde, J. & Rong, F. (2011): Sensorimotor integration in speech processing: Computational basis and neural organization. *Neuron* 69(3), 407-422.

Kang, E., Lee, D.S., Kang, H., Hwang, C.H., Oh, S.H., Kim, CS., Chung, J.K. & Lee, M.C. (2006): The neural correlates of cross-modal interaction in speech perception during a semantic decision task on sentences: a PET study**.** *Neuroimage* 32(1), 423-431.

Kovács, I., (2010): “Hot Spots” and Dynamic Coordination in Gestalt Perception. In: Malsburg, C., Philips A.W. & Singer, W. (eds.): *Dynamic Coordination in the Brain. From Neuront to Mind.* Cambridge, MA: The MIT Press, 215-227.

MacWhinney, B. (2005): Language evolution and human development. In: Bjorklund, D. & Pellegrini, A. (eds.): *Origins of the Social Mind: Evolutionary Psychology and Child Development*. New York: Guilford, 383-410.

McQueen, J., Cutler, A. & Norris, D. (2006): Phonological Abstraction in the Mental Lexicon. *Cognitive Science: A Multidisciplinary Journal* 30(6), 1113-1126.

Merabet, L.B. & Pascual-Leone, A. (2010): Neural reorganization following sensory loss: The opportunity of change. *Nature Reviews Neuroscience* 11(1), 44-52.

Mitterer, H., Chen, Y. & Zhou, X. (2011): Phonological abstraction in processing lexical-tone variation: evidence from a learning paradigm. *Cognitive Science* 35(1), 184-197.

Molnár, J. (1973): A magyar beszédhangok atlasza [The Atlas of Hungarian Speech Sounds]. Budapest: Tankönyvkiadó.

Von Muenster, K. & Baker, E. (2014): Oral communicating children using a cochlear implant: good reading outcomes are linked to better language and phonological processing abilities. *International Journal of Pediatric Otorhinolaryngology* 78(3), 433-444.

Pléh, Cs. (2006): A gyermeknyelv [Child language]. In: Kiefer, F. (ed.): *Magyar nyelv* [Hungarian Language]. Budapest: Akadémiai Kiadó, 2006.

Potamianos, G., Neti, C. & Deligne, S. (2003): *Joint Audio-Visual Speech Processing for Recognition and Enhancement.* AVSP 2003, Jorioz, France. Proc, 95-104.

Sjerps, M.J., & McQueen, J.M. (2010): The bounds on flexibility in speech perception. *Journal of Experimental Psychology: Human Perception and Performance* 36, 195-211.

Weisleder, P. (2012): No such thing as a “blind culture”. *Journal of Child Neurology* 27(6), 819-820.

Zahnert, T. (2011): The differential diagnosis of hearing loss. *Deutsches Arzteblatt International* 108(25), 433-443.

Zwicker, E. & Terhardt, E. (1980): Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. *Journal of the Acoustical Society of America* 68, 1523-1525.

László Hunyadi
University of Debrecen
H-4032 Debrcen, Egyetem tér 1.
hunyadi@unideb.hu

István Szekrényes
University of Debrecen
H-4032 Debrcen, Egyetem tér 1.
xepenator@gmail.com

László Czap
University of Miskolc
H-3515 Miskolc-Egyetemváros
czap@mazsola.iit.uni-miskolc.hu

István Sziklai
University of Debrecen
H-4032, Debrecen, Nagyerdei krt. 98.
isziklai@med.unideb.hu