PHONOLOGICAL ASPECTS OF AUDIOVISUAL SPEECH PERCEPTION

Ingo Hertrich, Werner Lutzenberger, Hermann Ackermann

University of Tuebingen
ingo.hertrich@uni-tuebingen.de

ABSTRACT

Based on magnetoencephalographic (MEG) measurements, this contribution tries to delineate a sequence of processing stages engaged in audiovisual (AV) speech perception, giving rise, finally, to the fusion of phonological features derived from auditory and visual input. Although the two channels interact even within early time windows such as the auditory M50 field, the definite percept appears to emerge at a relatively late stage (> 250 ms after the onset of the acoustic stimulus). Most noteworthy, the obtained data indicate visual motion to be encoded as categorical information even prior to AV fusion, as demonstrated by a non-linear visual /ta/ - /pa/ effect (within the time interval of the auditorily evoked M100 field) upon the strength of a magnetic source localized outside the auditory cortex. Taken together, these findings indicate, first, modality-specific sensory input to be transformed into phonetic features prior to the generation of a definite phonological percept and, second, cross-modal interactions to extend across a relatively large time window. Conceivably, these integration processes during speech perception are not only susceptible to visual input, but also to other supramodal influences such as top-down expectations and interactions with lexical data structures.

Keywords: Audiovisual speech perception, MEG, evoked magnetic fields, phonetic / phonological features, underspecification.

1. INTRODUCTION

1.1. Speech sounds in the brain

There is growing evidence that our brain maps incoming acoustic speech signals onto phonological representations at rather early stages of central-auditory processing. For example, electrophysiological studies indicate a fast-acting left-hemisphere mechanism for the detection of native vowel prototypes [15, 20]. As concerns dynamic components of the acoustic signal, furthermore, the brain seems to be "tuned" to events of a duration of ca. 40 ms, a time interval characterizing formant transitions in natural speech [10]. Psychoacoustic phenomena such as categorical perception or magnet effects suggest this ability of fast speech sound-related categorization to operate at the costs of the capability to discriminate subtle within-category differences. These mechanisms work in a native language-specific manner and, thus, seem to rely on "firm-wired" patterns, entrenched at early stages of language acquisition. Furthermore, fast categorical speech sound encoding represents a highly automatized process which does not depend upon attentional resources directed towards the acoustic channel.

It is well established that the human perceptual system integrates information across different sensory sources. Examples are "fusion errors", arising in dichotic listening experiments (e.g., left-ear /da/ concomitant with right-ear /pa/ resulting in perceived /ta/), cross-modal illusions such as the McGurk effect (see below), or the merging of lexical and bottom-up information [7]. The integration window for these interactions seems to extend across a quite large time interval (exceeding 150 ms), as indicated, for example, by the perceptual stability of cross-modal effects in case one stimulus channel is time-shifted against the other [14].

Considering the association of fast pre-attentive categorical stimulus encoding, on the one hand, with a broad cross-modal temporal tolerance, on the other, at least two subsequent stages of perceptual processing prior to the generation of a definite percept must be expected. In the following paragraphs, various aspects of AV speech perception will be addressed, providing further insights into the mechanisms of phonetic encoding.

1.2. Audiovisual Phonology

Particularly under noisy acoustic conditions, visual information may considerably improve speech perception [18]. The high efficiency of cross-modal information flow and the reproducibility of AV illusions such as the McGurk effect point at the operation of a highly automatized mechanism, working, presumably, at the level of economically stored information units such as phonetic or phonological features. The classical variant of the AV McGurk illusion is characterized by the auditory perception of the syllable /da/ in response to an acoustic /ba/ stimulus paired with a video displaying the production of /ga/. In terms of phonological features, the visual channel "cancels" the auditory feature /labial/ while the remaining sound characteristics (voiced stop) remain preserved. The visual feature /dorsal/ cannot be recognized, first, because it is not unambiguously signalized within the visual domain – at least for people untrained in speechreading – and, second, because it would, to some degree, contradict the information provided by the auditory channel, e.g., with respect to voice onset time and the intensity profile of the initial burst and interarticulator frication noise. Assuming that speech perception has to operate under time-critical forced-choice conditions, a decision has to be made on the basis of the available ambiguous or incomplete information. Some phonological models postulate an asymmetric specification of place of articulation and assign a higher value within the markedness hierarchy to the features /labial/ and /dorsal/, as compared to the underspecified cognate /coronal/. In case, labial and dorsal features are not unambiguously signalled, the underspecified item, as a consequence, will be set by default, giving rise to the percept of /da/.

visual /ga/	acoustic /ba/		perceived /da/
-labial	~~+labial~~	==>	-labial
~~+dorsal~~	-dorsal	==>	-dorsal

Figure 1: A possible phonological explanation of the McGurk effect: The acoustic "labial" and the visual "dorsal" (if present at all) features are neutralized by cross-modal interactions, giving rise to default setting of the underspecified feature "coronal".

1.3. Models of audiovisual fusion

Based on psychoacoustic findings, a variety of models have been proposed to explain AV fusion phenomena, differing in the time course of AV interactions and in the associated processing stages [17]. The central issue in this discussion is the question whether visual information is transformed into a phonological code prior to its fusion into an integrated auditory-phonetic percept or not. In particular, "analogue identification" models such as the suggestion of a "dominant recoding" of visual input within the auditory domain can be distinguished from "separate identification models", e.g., the "fuzzy logical theory of speech perception".

Electrophysiological methods such as electroencephalography (EEG) and magnetoencephalography (MEG) may shed, because of a high temporal resolution of the measured neural activity, further light on the cross-modal interactions bound to AV speech perception. Among others, the available EEG and MEG studies revealed an impact of visual speech upon auditorily evoked activity at different processing stages:

- visually-induced suppression of the auditory P50 potential, i.e., a characteristic response to any acoustic signal, peaking about 50 ms after stimulus onset [12],
- visual enhancement of subcomponents of the auditory N1 wave [8], i.e., an evoked response with a latency of ca. 100 ms,
- N1 attenuation effects [19],
- cross-modal hypoadditive event-related interactions at a delay of 120 - 190 ms after the onset of the acoustic signal, and
- visually induced mismatch negativities (MMN) or mismatch fields (MMF) in response to the 'deviant' stimuli of an 'oddball' design, (latency = 150-300 ms, [12, 13, 4,5]).

A recent study of our group [11] found, based upon an oddball-design, at least two different time windows of AV interactions. (a) The phonetic fusion of auditory and visual information seems to occur at a quite late stage of sensory memory processing, as indicated by a speech-specific visually-induced left-lateralized component of the mismatch field at a latency of about 275 ms after the onset of the acoustic stimulus. (b) By contrast, earlier effects, acting upon M50 and M100 deflections, were also observed when visual speech was paired with tone signals and, thus, seem to be bound to speech-unrelated attentional processes. Several animals studies provided some further evidence for an early impact of visual information upon auditorily evoked cortical activity [2, 3].

In the following section, a MEG experiment will be discussed, comparing visual speech and non-speech motion effects on evoked magnetic fields within the time domain of the auditory M50 and M100 components.

2. EARLY AV EFFECTS DURING SPEECH PERCEPTION: A MEG EXPERIMENT

In order to evaluate speech-specific early visual influences upon auditorily evoked responses, an MEG experiment was performed, including all four combinations of acoustic and visual speech and non-speech conditions. Since the experiment, furthermore, encompassed visual-only and acoustic-only control stimuli, AV interactions could be characterized in terms of hypo- or hyperadditive effects.

2.1. Methods

25 healthy right-handed subjects participated in the MEG experient, anatomical MRI datasets could be obtained from 17 participants.

Four different AV configurations (acoustic / visual sensory modality × speech / non-speech events) were implemented in different runs of the experiment. Figure 1 displays the temporal structure of the AV speech stimuli.An ambiguous CV syllable generated by means of a formant synthesizer [9] served as the acoustic speech stimulus. It could be perceived either as /pa/ or /ta/, depending on a synchronized video, displaying either /pa/ or /ta/ articulation. In other words: visual information disambiguated the acoustic signal. Fundamental frequency (F0) amounted to 120 Hz across the initial part (duration = 200 ms) of the stimulus. Following this stationary phase, F0 began to rise or to fall by six semitones to either 170 or 85 Hz, respectively, at signal offset (300 ms). This F0 movement was introduced to direct subjects' attention toward the auditory channel, using a pitch recognition task. Since the present MEG measurements were restricted to a time interval preceding the onset of the pitch movements, the upward or downward F0 shift had no direct impact on the MEG data.

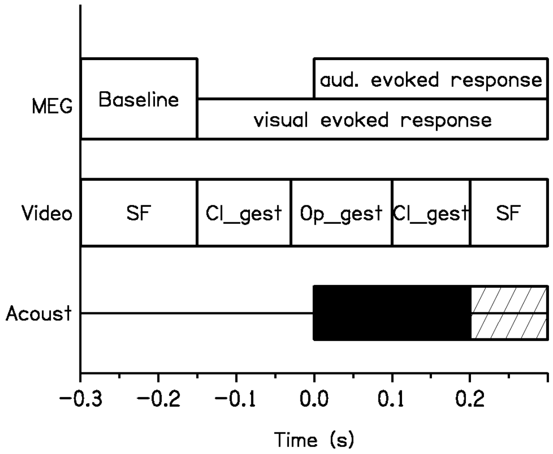

Figure 2: Time course of a single AV speech trial. Acoust: duration of the acoustic signal, the hatched part corresponds to the final pitch movement. Video: SF = static face, continuously displayed between the stimuli, Cl_gest / Op_gest = duration of the visible mouth closing and opening gestures. MEG: baseline = pre-stimulus interval serving as the baseline of the MEG sensor data. The onset of visual motion precedes the acoustic signal by 150 ms.

A periodic signal consisting of repetitions of single-formant sweeps served as the acoustic non-speech stimulus. Within each single pitch period, a formant was down-tuned from 2000 to 500 Hz and dampened to zero at its offset. The periodic structure of these signals gives rise to a strong pitch percept, lacking, however, any resemblance to speech sounds. Similar to the speech stimuli, F0 amounted to 120 Hz during the first 200 ms, followed by an upward or downward pitch movement.

The visual speech condition comprised two different videos, showing a male speaker uttering the syllable /pa/ or /ta/, respectively. These sequences of a duration of 300 ms each were embedded into a larger frame, extending across a time interval of 1.4 s (= onset-to-onset inter-stimulus interval). In other words: A static display of the same speaker's face preceded and followed the visual /pa/ and /ta/ sequences. As a consequence, the visual speech stimuli could be concatenated into larger runs, in the absence of any noticeable discontinuities of the video display.

During the visual non-speech condition, contraction / expansion of concentric circles served as an analogue to orofacial motion associated with the /pa/ and /ta/ utterances. Two different movement sequences of a duration of 300 ms each were created (contraction and expansion time = 150 ms each), i.e., larger/faster and smaller/slower excursions, in analogy to the kinematic characteristics of visible lip articulation bound to /pa/ and /ta/ syllables. Again, the same static picture preceded and followed the circle movements.

The four AV configurations referred to were implemented in different runs. Six relevant stimulus categories (three levels of movement: large, small, or no motion, in the presence or absence of an acoustic signal) were applied in a pseudo-randomized order in each run:

1. large motion (or /pa/) - with acoustic signal
2. large motion (or /pa/) - no acoustic signal
3. small motion (or /ta/) - with acoustic signal
4. small motion (or /ta/) - no acoustic signal
5. static picture - with acoustic signal
6. static picture - no acoustic signal

The silent visual and the static acoustic stimuli had been added (visual-only and acoustic-only conditions) in order to separate hypo- or hyperadditive cross-modal effects of the two sensory modalities. The "empty" stimulus (no. 6) served as a control, allowing for the detection of unspecific expectation effects due to the regular inter-stimulus intervals.

Using a whole-head device (CTF, Vancouver, Canada; 151 sensors, sampling rate = 312.5 Hz, anti-aliasing filter cutoff = 120 Hz), evoked magnetic fields were recorded across a time-interval of 550 ms, starting 150 ms prior to the onset of orofacial speech movements or non-speech motion of the video display. The initial interval of 150 ms served as the pre-stimulus baseline. MEG offset was removed from each sensor signal by subtracting its mean baseline value. An automatic software procedure allowed for the detection of eyeblink artefacts, and the respective trials (ca. 5-10%) were discarded from analysis.

Anatomical MRI datasets were transformed into the head-related coordinates of the MEG device (orthogonal axes based on the two pre-auricular points and the nasion, resolution = 1 mm, 256 × 256 × 256 matrix). As a head model for MEG dipole analysis, based on the group data, an MRI image across all 17 available datasets was created (voxel-wise averaging and gray-scale normalization to the dynamic range of the display program; 'MRIViewer', CTF, Vancouver). In spite of individual variability of head size and shape, the MRI group average displayed all the relevant brain structures.

2.2. Results and discussion

2.2.1. Dipole model

A 6 dipole-model was used to delineate the time course of evoked brain activity, comprising 3 bilateral sources:

1. a pair of dipoles localized within auditory cortex (A110),
2. a pair of sources bound to visual cortex, modelling motion-induced activity at a latency of 170 ms after the onset of visual motion (V170), and
3. a pair of dipoles within the posterior insula, corresponding to a peak of visually-induced activity at 270 ms (V270; temporal overlap with auditory M100, see Figure 3).

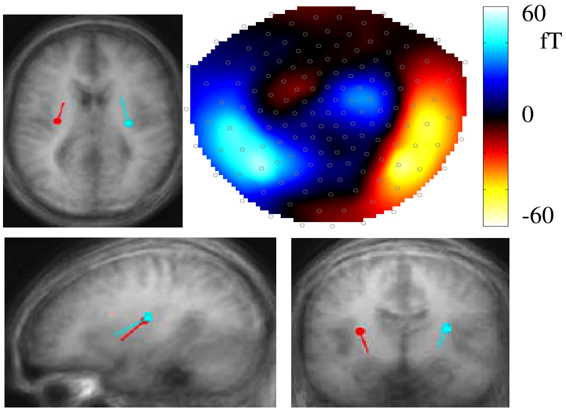

Figure 3: MEG brain map (upper right panel) of the V270 field (270 ms after motion onset), concomitant with the anatomical location of the dipole sources (data based on group averages). MRI slices correspond to the left-hemisphere dipoles.

2.2.2. Visual impact on auditorily evoked M50

Figure 4 displays the grand average of the time course of the responses to AV simuli. Within the domain of the auditory M50 field (50-70 ms after acoustic stimulus onset), significant visual effects were observed in subspace projections onto all three dipole pairs. The auditory dipole shows a visually-induced attenuation, in line with a previous study, reporting visually-induced M50 suppression (main effect of motion on A110 dipole moment: $F[1,24] = 19.94$, $p < 0.001$). Since a similar effect emerged during application of visual-only stimuli, these findings, conceivably, reflect a "preparatory baseline shift" of the central-auditory system [6].

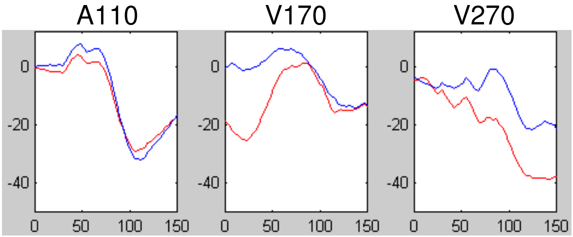

Figure 4: Time course (from acoustic onset onwards, in ms) of subspace projections (nAm) onto the A110, the V170, and the V270 dipoles (pooled across both hemispheres): Effects of large (or /pa/) visual motion (red) in comparison to the static conditions (blue), pooled across the speech and non-speech AV conditions.

2.2.3. *Visual impact on auditorily evoked M100*

In contrast to the M50 component, the speech and non-speech conditions were characterized by different effects within the M100 time window. Non-speech motion resulted in an attenuation of the A110 dipole source (upper left panel of Figure 5), depending, however, on the presence of auditory input (interaction of visual motion with the presence or absence of an acoustic signal: $F[1,24] = 5.10$, $p < 0.05$), whereas significant influences of speech gestures upon the A110 dipole strength were restricted to silent stimuli, indicating hypoadditive cross-modal effects (interaction of visual motion with the presence or absence of an acoustic signal: $F[1,24] = 5.67$, $p < 0.05$). The subspace projections onto the A110 dipole source failed to exhibit any relevant non-linear impact of visual motion, i.e., the responses to small motion (or /ta/) did not significantly differ from an intermediate response between large motion and the respective static display.

The V270 dipole source showed similar motion effects for visual /pa/ and large non-speech excursions (red lines in the right panels of Figure 5). Furthermore, both responses to the small movements (small non-speech excursions and /ta/ syllable, respectively) differed from an intermediate response (nonlinear motion effect /ta/: $F[1,24] = 4.40$, $p < 0.05$; non-speech: $F[1,24] = 6.87$, $p < 0.05$). Additionally, the /ta/ effect showed a 3-way interaction with hemisphere and the presence or absence of an acoustic signal $F[1,24] = 4.82$, $p < 0.05$). Post hoc analysis revealed the left-hemisphere response to visual /ta/ to be particularly suppressed in the absence of an acoustic signal. At this stage of processing, visual motion, thus, showed an all-or-nothing effect, eventually indicating the "markedness" of the visual event. In case of visual non-speech stimuli, by contrast, both the small and the large movements acted as a pre-cue of the following acoustic event. Conceivably, the large lip movements associated with /pa/, reflecting a marked phonetic feature, served as an above-threshold signal whereas the small /ta/ excursions, signalling an underspecified place of articulation [1], were suppressed.

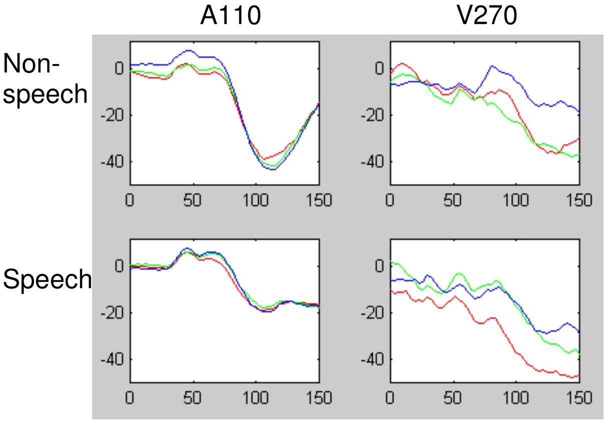

Figure 5: Time course of subspace projections onto the A110 (left panels) and the V270 dipoles (right panels) of the 6 dipole-model: Effect of large (red) and small (green) movement excursions in comparison to the static (blue) conditions (upper panels = AV non-speech, lower panels = AV speech).

3. CONCLUSIONS

The observed motion effects suggest the visual channel to influence the central-auditory system at several different processing stages. The red arrows in Figure 6, indicating statistically significant influences of visual speech and non-speech events upon the M50 field, the M100 component, and the mismatch field (MMF), refer to three subsequent cross-modal interactions.

Since visual cues preceded the acoustic signal by ca. 150 ms, the early visual impact upon the M50 field, observed both in response to speech and non-speech motion, presumably, represents a visually-induced baseline shift or pre-activation of the central-auditory system, as documented, e.g., in animal experiments [2].

The M100 field has been assumed to reflect pre-representational processes, related to the detection of specific signal properties [16]. At this stage, the speech and non-speech conditions yielded a different pattern of AV interactions. Within this time interval, the specific phonetic

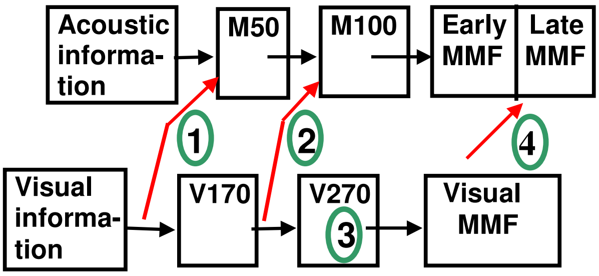

Figure 6: Hypothesized sequence of AV interactions (the numbers refer to distinct effects of visual motion on evoked magnetic fields):
(1) Unspecific M50 attenuation (preparatory baseline shift).
(2) AV interactions within the domain of the M100 component, differential impact of visual speech and non-speech information (speech: hypoadditive enhancement; non-speech: attenuation).
(3) Phonological weighting of visual input outside the central-auditory system: left-hemisphere suppression of the phonologically unmarked visual /ta/-gesture.
(4) Cross-modal sensory memory operations, giving rise to a fused phonetic precept, as indicated by a speech-specific visually-induced left-lateralized late (275 ms) MMF component [11].

visual information, i.e., the presence or absence of the labial phonological feature, seems still to be represented outside the cortical auditory system, being already transformed, however, into a binary code, as indicated by the non-linear visual /pa/-/ta/ effect. By contrast, a previous oddball experiment suggested the actual fusion of phonetic information into an integrated auditory percept to be bound to a later phase of sensory memory processing [11].

Taken together, thus, these MEG experiments were able to separate earlier unspecific from later speech-related cross-modal AV interactions. Most noteworthy, sensory information appears to be transformed into phonetic features (binary-coded information) prior to the emergence of a definite phonological percept at a later processing stage.

4. REFERENCES

[1] Avery, P., Rice, K. 1989. Segment structure and coronal underspecification. *Phonology* 76, 179-200.

[2] Bizley, J.K., Nodal, F.R., Bajo, V.M., Nelken, I., King, A.J. 2006. Physiological and anatomical evidence for multisensory interactions in auditory cortex. *Cereb. Cortex*, doi:10.1093/cercor/bhl128.

[3] Brosch, M., Selezneva, E., Scheich, H. 2005. Nonauditory events of a behavioral procedure activate auditory cortex of highly trained monkeys. *J. Neurosci.* 25, 6797-6806.

[4] Colin, C., Radeau, M., Soquet, A., Deltenre, P. 2004. Generalization of the generation of an MMN by illusory McGurk percepts: voiceless consonants. *Clin. Neurophysiol.* 115, 1989-2000.

[5] Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., Deltenre, P. 2002. Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. *Clin. Neurophysiol.* 113, 495-506.

[6] Driver, J., Frith, C. 2000. Shifting baseline in attention research. *Nature Rev. Neurosci. 1,* 147-148.

[7] Ganong, W.F. 1980. Phonetic categorization in auditory word perception. *J. Exp. Psychol.: Hum. Perc. Perf.* 6, 110-125.

[8] Giard, M.H., Peronnet, F. 1999. Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *J. Cogn. Neurosci.* 11, 473-490.

[9] Hertrich, I., Ackermann, H. 1999. A vowel synthesizer based on formant sinusoids modulated by fundamental frequency. *J. Acoust. Soc. Am.* 106, 2988-2990.

[10] Hertrich, I., Mathiak, K., Lutzenberger, W., Ackermann, H. 2003. Processing of dynamic aspects of speech and non-speech stimuli: a whole-head magnetoencephalography study. *Cogn. Brain Res.* 17, 130-139

[11] Hertrich, I., Mathiak, K., Lutzenberger, W., Menning, H., Ackermann, H. 2007. Sequential audiovisual interactions during speech perception: a whole-head MEG study. *Neuropsychologia* 45, 1342-1354.

[12] Lebib, R., Papo, D., De Bode, S., Baudonniere, P.M. 2003. Evidence of a visual-to-auditory cross-modal sensory gating phenomenon as reflected by the human P50 event-related brain potential modulation. *Neurosci. Lett.* 341, 185-188.

[13] Möttönen, R., Krause, C.M., Tiippana, K., Sams, M. 2002. Processing of changes in visual speech in the human auditory cortex. *Cogn. Brain Res.* 13, 417-425.

[14] Munhall, K.G., Gribble, P., Sacco, L., Ward, M. 1996. Temporal constraints on the McGurk effect. *Percept. Psychophys.* 58, 351-362.

[15] Näätänen, R., Lehtokoski, A., Lennes, M., Cheour, M., Huotilainen, M., Iivonen, A., Vainio, M., Alku, P., Ilmoniemi, R.J., Luuk, A., Allik, J., Sinkkonen, J., Alho, K. 1997. Language-specific phoneme representations revealed by electric and magnetic brain responses. *Nature* 385, 432-434.

[16] Näätänen, R., Winkler, I. 1999. The concept of auditory stimulus representation in cognitive neuroscience. *Psychol. Bull.* 125, 826-859.

[17] Schwartz, J.L., Robert-Ribes, J., Escudier, P. 1998. Ten years after Summerfield: a taxonomy of models for audio-visual fusion in speech perception. In: Campbell, R., Dodd, B., Burnam, D. (Eds.), *Hearing by Eye II.* Hove: Psychology Press, 85-108.

[18] Sumby, W.H., Pollack, I. 1954. Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212-215.

[19] Van Wassenhove, V., Grant, K.W., Poeppel, D. 2005. Visual speech speeds up the neural processing of auditory speech. *Proc.Natl.Acad. Sci.USA* 102, 1181-1186.

[20] Winkler, I., Lehtokoski, A., Alku, P., Vainio, M., Czigler, I., Csepe, V., Aaltonen, O., Raimo, I., Alho, K., Lang, H., Iivonen, A., Näätanen, R. 1999. Pre-attentive detection of vowel contrasts utilizes both phonetic and auditory memory representations. *Cogn. Brain Res.* 7, 357-369.