PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Audiovisual speech perception in children with autism spectrum disorders

Julia IRWIN[1,3]; Trey AVERY[3,4]; Daniel KLEINMAN[3]; Nicole LANDI[1,2],

[1] Southern Connecticut State University, USA

[2] University of Connecticut, USA

[3] Haskins Laboratories, USA

[4] Philips Healthcare, Netherlands

ABSTRACT

In face-to-face conversation, when a speaker talks, the outcome of their speech can both be heard (audio) and seen (visual). We employed a novel visual phonemic restoration paradigm to assess neural signatures (event related potentials [ERPs]) of audiovisual processing in typically developing children and in children with ASD. During EEG recording, two types of auditory stimuli were alternately presented with video of a speaker saying the consonant-vowel syllable /ba/: 1) a synthesized consonant-vowel syllable /ba/ or 2) a synthesized syllable derived from /ba/ in which auditory cues for the consonant are substantially weakened, such that it sounds more like /a/. The auditory stimuli are easily discriminable, however, in the context of a visual /ba/, the auditory /a/ is typically perceived as /ba/, producing a visual phonemic restoration. In an ERP context, we have shown that this restoration leads to an attenuated phoneme discrimination response in an active task in typical adults and children. To explore the hypothesis that children with autism spectrum disorder (ASD) have atypical AV speech integration under pre-attentive processing conditions, we tested whether children with ASD would show a reduction in this restoration effect under passive listening conditions. Indeed, in this task, children with ASD showed a large /ba/-/a/ discrimination response, even in the presence of a speaker producing /ba/, suggesting reduced influence of visual speech.

Keywords: Audiovisual, Speech, Autism, Phonemic Restoration

1. INTRODUCTION

In face-to-face interactions, listeners can both hear and see what a speaker is saying. Visual information about speech has been shown to influence what listeners hear, increasing identification of the speech signal both in noisy environments [1] and in clear listening conditions, where mismatched auditory and visual speech results in a new percept (known as the McGurk effect [2]). Natural listening environments are often noisy, particularly for children (e.g. classrooms, cafeterias, playgrounds), making the ability to use the multimodal speech signal central in understanding of the spoken message. Critically, any loss in the ability to understand a speaker's message can lead to cascading negative effects in social communication, already primary deficits for children with an autism spectrum disorder (ASD).

Children with an ASD have been reported to have impaired processing of audiovisual or (AV) speech in comparison to their typically developing (TD) peers, with data from a number of studies suggesting weaker integration of the face and voice [3-7]. These deficits in integration are indicated by less influence of the speaker's face in audiovisual tasks such as speech in noise (where the listener must integrate the face and degraded speech sound to identify what was said) and mismatched audiovisual speech (where the face and voice are different signals, and, if integrated, will lead to a visually influenced percept), both of which occur in a robust manner in children with TD. Weakened integration would significantly impair a listeners' ability to recover or disambiguate a speaker's message in noisy listening environments, which may account for some of the observed language difficulties in children with ASD.

In addition to these behavioral findings, a few recent studies have utilized event related potentials (ERPs), which provide an objective neural index of perception, to explore AV speech processing in individuals with ASD. For example, Magnée et al., 2008 [8] found that adults with ASD do not show typical congruency associated N2 effects (e.g. to a face and voice with matching vs. mismatching emotional valence), suggesting that adults with ASD are less sensitive to mismatching AV speech stimuli. Moreover, a recent ERP study found that infant siblings of individuals diagnosed with ASD (who are at greater risk of ASD themselves) show a prolonged latency in later P400 components in responding to direct gaze during face processing [9]. Taken together, these studies indicate that individuals with ASD have difficulty using visual speech information during perception of a speaking face. This does not appear to be simply due to less looking to the face of a speaker – Irwin, Tornatore, Brancazio & Whalen (2011) [6] controlled for this by examining AV speech perception only when children were fixated on the face of the speaker and reported significant differences in children with ASD as compared to TD controls in visual influence on heard speech.

While existing studies are suggestive of atypical response to AV speech in children with ASD and their siblings, the paradigms used to date for studying AV speech may not be ideal for use in this population because they make additional processing demands beyond perception of audio and visual speech. Specifically, studies that use speech in noise and/or mismatched (or McGurk type) AV tasks have potential limitations for young children and those with ASD [10]. The McGurk Effect creates a percept that differs from either the visual or auditory signal alone because of conflict between the two modalities. These percepts are identified as poorer exemplars of a category than matched A + V stimuli [11]. This approach may be particularly problematic for those with ASD because weaknesses in executive function can lead to difficulties in identification of ambiguous stimuli [12]. Additionally, studying AV speech perception using paradigms that utilize auditory noise is problematic because noise is generally disruptive for individuals with ASD in the perception of speech [13].

In order to examine visual influence on heard speech in children with autism, we have developed a measure that involves neither noise nor auditory and visual category conflict that can serve as an alternative to assessing audiovisual speech processing (also see [14] for a related approach). This new paradigm, which we describe in detail in Irwin et al. [15], uses restoration of weakened auditory tokens with visual stimuli. Two types of stimuli are presented to the listener: clear exemplars of an auditory consonant-vowel syllable (in this case, /ba/), and syllables in which the auditory cues for the consonant are substantially weakened, creating a stimulus which is more /a/ like, from this point on referred to as /*a*/. The auditory stimuli are created by synthesizing speech based on a natural production of the syllable and systematically flattening the formant transitions to create the /*a*/. Video of the speaker's face does not change (always producing /ba/), but the auditory stimuli (/ba/ or /*a*/) vary. Thus, when the /*a*/ auditory stimulus is dubbed over the visual /ba/, a visual influence will result in effectively "restoring" the weakened auditory cues so that the stimulus is perceived as a /ba/, akin to a visual phonemic restoration effect [16 - 18].

To provide a sensitive measure of AV speech processing in typically developing children and those with ASD, we measured neural signatures of audiovisual perception and integration using this novel visual phonemic restoration method. We utilize an equiprobable paradigm to elicit ERP responses to /*a*/ and /ba/. All speech tokens are paired with a face producing /ba/. In this paradigm, if the visual /ba/ causes the auditory /*a*/ to be perceived as /ba/ (phonemic restoration), then ERP response to both tokens should be similar. However, if audio and visual speech are not integrated (no phonemic restoration) we would expect a differential, or phoneme mismatch response. Here we hypothesize that children with ASD, who are suspected to exhibit deficits in AV integration, will be less likely to use visual speech to effectively restore the /*a*/ sound to a /ba/ percept. If this is the case, children with ASD would exhibit a larger mismatch response (here, mismatch negativity (MMN)) for the auditory /*a*/ - visual /ba/ condition relative to their typically developing (TD) controls.

2. Method

2.1 Participants

Fifty-one monolingual American English-speaking children participated: 18 children with ASD (13 males and 5 females, ages 7.6-14.9, mean age 10.7, SD 2.4) and 32 TD children (14 males and 18 females, ages 5.3-13.2, mean age 9.9, SD 2.0). All participants were right-handed, and had normal vision and hearing. Children in no history or developmental delays per parent report. For characterization purposes, all participants in the ASD group had a clinical diagnosis of an ASD and completed the Autism Diagnostic Observation Schedule – second edition [ADOS-2, 19] and their parents completed the Autism Diagnosis

Interview Revised [ADI-R, 20]. Performance IQ subtests from the Wechsler Abbreviated Scale of Intelligence were used for matching of groups [21]. All data were collected according to the ethical guidelines laid out by the Yale University Institutional Review Board. Written consent was obtained from participants' primary caregivers and written assent from child participants.

Due to differences between samples in demographic characteristics, we report two sets of analyses below. The first analysis includes data from all participants in both groups. The second analysis includes data from a sample of 15 participants in each group who were matched on sex, age, composite WASI scores, and number of usable trials (the average across conditions) using the program Match [22]. Each sample included 10 males and 5 females, and samples did not significantly differ on any of the characteristics on which participants were matched: age (ASD: Range = 7.6-14.9, M = 11.0, SD = 2.4; TD: Range = 6.7-13.2, M = 10.6, SD = 2.0; $t(28) = -0.46$, $p = .65$), composite WASI score (ASD: M = 99.5, SD = 17.1; TD: M = 105.5, SD = 10.9; $t(28) = 1.15$, $p = .26$); or average number of usable trials per condition (ASD: M = 45.3, SD = 19.4; TD: M = 49.5, SD = 18.0; $t(28) = 0.61$, $p = .55$).

2.3 Audiovisual Stimuli and Experimental Paradigm

Participants completed three short audiovisual (AV) experiments designed to examine the neural basis of AV speech integration in our participants. The primary experiment of interest (the AV speech experiment) utilized the novel phonemic restoration procedure introduced above and will be reported here. This experiment included video of a male speaker who produced the syllable /ba/ while participants were presented with either a full /ba/ or the reduced /ba/ (or /*a*/).

To produce the stimuli for the AV speech experiment, the /ba/ and /*a*/ synthesized auditory stimuli were dubbed onto video of the speaker producing /ba/, with the acoustic onsets synchronized with the visible articulation time locked to a single video frame. The stimuli for the AV speech were created by videotaping and recording an adult male speaker of English producing the syllable /ba/. Using Praat, we extracted acoustic parameters for the token, including formant trajectories, amplitude contour, voicing and pitch contour [23]. Critically, the token had rising formant transitions for F1, F2, and to a lesser extent F3, characteristic of /ba/. To create our /ba/ stimulus, we synthesized a new token of /ba/ based on these values. To create our /*a*/ stimulus we then modified the synthesis parameters by changing the onset values for F1 and F2 to reduce the extent of the transitions and lengthened the transition durations for F1, F2 and F3, and then synthesized a new stimulus.

Instructions and a practice trial were presented prior to the start of the electroencephalography (EEG) session, which consisted of 3 experiments containing 200 equiprobable pseudorandomized presentation of 100 /ba/ and 100 /*a*/ tokens each. Total EEG session time was approximately 45 minutes, depending on length of breaks and amount of EEG net rehydration between experiments.

EEG Data Collection

EEG data was collected with a Philips Neuro (formerly Electrical Geodesics Inc) EEG System using 128 Ag/AgCl electrodes embedded in soft sponges woven into a geodesic array. The EEG sensor nets were soaked for up to ten minutes prior to use in a warm potassium chloride solution. Impedance for all electrodes was kept below 40 kΩ throughout the experimental run (impedances were re-checked between experiments – approximately every 15 minutes). Online recordings at each electrode used the vertex electrode as the reference and were later referenced to the average reference. EEG was continuously recorded using Net Station 4.5 on a MacPro. Stimuli were presented using E-Prime version 2.0.8.90 (Psychology Software Tools, Inc., Sharpsburg, PA, USA) on a Dell computer running Windows XP. Audio stimuli were presented from an audio speaker centered 85 cm above the participant at 65 decibels. Visual stimuli were presented on a Dell 17-inch flat panel monitors 60 cm from the participant.

ERP Data Preprocessing

Preprocessing was conducted using Net Station 5.4. EEG data were band-pass filtered between .3 and 30 Hz (Passband Gain: 99.0% [−0.1 dB], Stopband Gain: 1.0% [−40.0 dB], Rolloff: 2.00 Hz) and segmented by condition, 100 ms pre-stimulus to 800 ms post-stimulus. Eye blinks and vertical eye movements were examined with electrodes located below and above the eyes (channels 8, 126, 25, 127). Horizontal eye movements were measured using channels 125 and 128, located at positions to the left and right of the eyes. Artifacts were automatically detected and manually verified for exclusion from additional analysis (bad channel >200 microvolts, eye blinks >140 microvolts and eye movement >55 microvolts). For every channel,

33% or greater bad segments was used as the criteria for marking the channel bad; for every segment, greater than 20 bad channels was used as a criterion for marking a segment bad. Participants with less than 15% of a possible 100 usable trials in any condition were excluded from analysis. The average usable trial count across all conditions was a mean of 51.0 (SD = 19.6) and each condition had similar amounts of usable data, /ba/ mean = 50.9 (SD = 20.5) and /a/ mean = 51.1 (SD = 19.5). Bad channels (fluctuations over 200 μV) were spherical spline interpolated from nearby electrodes. Data were baseline-corrected using a 100 ms window prior to onset of all stimuli. Data were re-referenced from vertex recording to an average reference of all 129 channels. All processed, artifact-free segments were averaged by condition, producing a single event-related potential waveform for each condition.

ERP Data Analysis

Electrode montages and temporal windows were selected based on a combination of previous literature, including our own findings in a similar paradigm, and visual inspection [26, 27]. Visual inspection of our data revealed two negative mismatch associated peaks within the expected central scalp location; the first fell in between 250 and 400 milliseconds (Early Mismatch Negativity; Early MMN), and the second fell between 400 and 550 milliseconds (Late Mismatch Negativity; Late MMN). For analysis, we identified the most negative peak in a cluster of eleven central electrodes (see Figure 2) within each latency window for each individual and used the amplitude of this peak as the dependent variable. To examine the ERP effects of speech stimulus (congruent /ba/ vs. incongruent /*a*/) by group, we ran a 2 × 2 × 2 repeated measures ANOVAs with group (ASD vs. TD) as a between-subjects factor, and latency window (Early MMN vs. Late MMN) and speech stimulus (congruent vs. incongruent) as within-subjects factors. To examine the ERP effects of speech stimulus (congruent /ba/ vs. incongruent /*a*/) by group, we ran a 2 × 2 × 2 repeated measures ANOVAs with group (ASD vs. TD) as a between-subjects factor, and latency window (Early MMN vs. Late MMN) and speech stimulus (congruent vs. incongruent) as within-subjects factors.

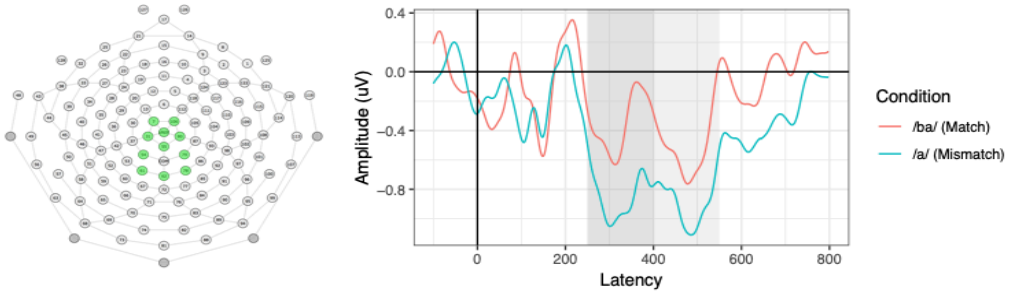
3. RESULTS

All participants. No main effects or two-way interactions reached statistical significance, all Fs < 2.2, all ps > .14. Crucially, however, the three-way interaction between group, latency window, and speech stimulus was statistically significant, $F(1,48) = 5.29$, $p = .026$, indicating different patterns of sensitivity to AV-congruent and AV-incongruent stimuli across groups in the two latency windows.

To explore this interaction further, we conducted planned contrasts separately for each group and latency window. In the Early MMN window, the ASD group showed a greater negativity to incongruent stimuli than to congruent stimuli, $B = -1.13$ μV, 95% CI = ±0.48 μV, $t(48) = -4.78$, $p < .0001$, but the TD group did not, $B = -0.01$ μV, 95% CI = ±0.35 μV, $|t| < 1$. Consistent with this pattern, the ASD group showed a larger negativity to AV-incongruent stimuli than the TD group in the Early MMN window, $B = -0.56$, 95% CI = ±0.30 μV, $t(48) = -3.78$, $p < .0004$.

In the Late MMN latency window, neither group showed sensitivity to AV-congruence, ASD: $B = -0.36$ μV, 95% CI = ±0.48 μV, $t(48) = -1.54$, $p = .13$; TD: $B = -0.21$, 95% CI = ±0.35 μV, $t(48) = -1.16$, $p = .25$, and groups did not differ in this (lack of) sensitivity, $B = -0.08$, 95% CI = ±0.30 μV, $|t| < 1$.

Fig.1 (left panel). The electrode montage. Electrodes included in the cluster for analysis (n=11) are colored in green; Fig. 2. (Right panel). Grand mean ERP waveforms for all groups, for the electrode cluster shown in Fig. 2. Early and late MMN windows are highlighted in darker and lighter grey boxes, respectively.

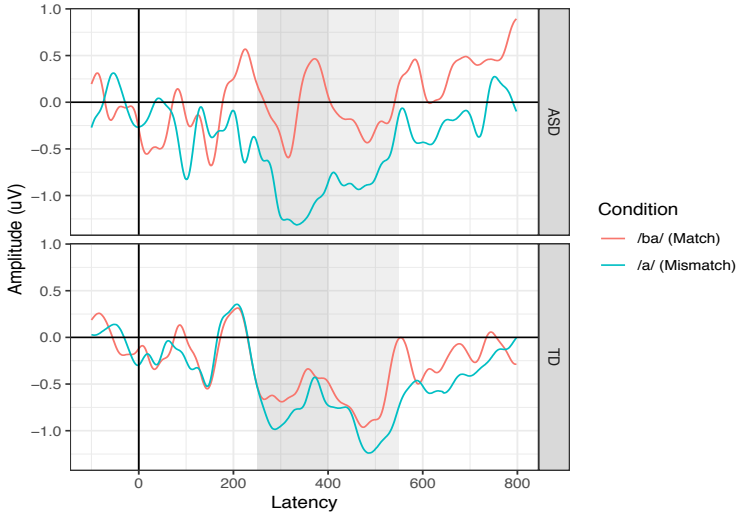


Matched participant samples. No main effects or two-way interactions reached significance, all *F*s < 1.3, all *p*s > .26. The three-way interaction between group, latency window, and speech stimulus did not reach significance either, $F(1, 28) = 2.88$, $p = .101$.

We conducted planned contrasts separately for each group and latency window, which revealed a pattern that was statistically identical to the analysis of all participant data. In the Early MMN window, the ASD group showed a greater negativity to incongruent stimuli than to congruent stimuli, $B = -0.94$ µV, 95% CI = ±0.61 µV, $t(28) = -3.15$, $p = .004$, but the TD group did not, $B = 0.33$, 95% CI = ±0.61 µV, $t(28) = 1.10$, $p = .28$. Consistent with this pattern, the ASD group showed a larger negativity to AV-incongruent stimuli than the TD group in the Early MMN window, $B = -0.63$, 95% CI = ±0.43 µV, $t(28) = -3.00$, $p = .006$.

In the Late MMN latency window, neither group showed sensitivity to AV-congruence, ASD: $B = -0.24$ µV, 95% CI = ±0.61 µV, $|t| < 1$; TD: $B = 0.01$, 95% CI = ±0.61 µV, $|t| < 1$, and groups did not differ in this (lack of) sensitivity, $B = -0.13$, 95% CI = ±0.43 µV, $|t| < 1$.

Fig. 3. Grand mean ERP waveforms separately for each group, for the electrode cluster shown in Early and late MMN windows are highlighted in darker and lighter grey boxes, respectively.



4. CONCLUSIONS

Children with an autism spectrum disorder have been reported to be less influenced by visible speech even when controlling for gaze to the face of a speaker. However, the underlying mechanisms of deficits in audiovisual speech perception are still unknown. We employed a novel visual phonemic restoration paradigm in a passive ERP paradigm to assess neural signatures of audiovisual processing in typically developing children and in children with ASD. The two stimulus types, an auditory consonant-vowel syllable /ba/ and a

syllable in which the auditory cues for the consonant was substantially weakened (/*a*/) were paired with video of a speaker producing /ba/. In this paradigm, the video of the speaker producing /ba/ can restore perception of the consonant when paired with the auditory /*a*/, leading to a perception of /ba/. If this phonemic restoration occurs, it should lead to an attenuated ERP discrimination response (here, the MMN), providing a measure of integration of the visual and auditory speech signals. We found that children with ASD had a larger MMN response relative TD children, suggesting reduced audiovisual integration. We note however that one limitation of the current study is that we did not have an auditory only control condition for exploring the /ba/ vs. /a/ contrast with no visible speech, thus it is unknown whether children with ASD have atypical patterns of speech perception for this contrast.

.

These findings provide preliminary evidence of impaired integration of auditory and visual speech signals in ASD in a passive AV speech processing paradigm. Because children with ASD showed a greater /ba/-/a/ discrimination response in the presence of video of the speaker producing /ba/ in this passive condition, this indicates that children with ASD show reduced influence of visual speech. Future work will directly compare passive and active ERP response to AV speech and will include an auditory only condition.

5. ACKNOWLEDGEMENTS

Supported by NIH grants DC013864 and DC011342.

6. REFERENCES

[1.] Sumby, W. H.; Pollack, I. Visual contribution to speech intelligibility in noise. J. Acoust. Soc. Am., 1954; Volume 26(2), 212-215.

[2.] McGurk, H., & MacDonald, J. Hearing lips and seeing voices. Nature, 1976; 746-748.

[3.] Foxe, J.J.; Molholm, S.; Del Bene, V.A.; Frey, H.P.; Russo, N.N.; Blanco, D.; Ross, L.A. Severe multisensory speech integration deficits in high-functioning school-aged children with autism spectrum disorder (ASD) and their resolution during early adolescence. Cereb. Cortex, 2015; 25, 298–312.

[4.] Iarocci, G., Rombough, A., Yager, J., Weeks, D. J., & Chua, R. Visual influences on speech perception in children with autism. Autism, 2010; 14(4), 305-320.

[5.] Irwin, J. R., Tornatore, L. A., Brancazio, L., & Whalen, D. H. Can children with autism spectrum disorders "hear" a speaking face? Child Dev, 2011; 82(5), 1397-1403.

[6.] Mongillo, E. A., Irwin, J. R., Whalen, D. H., Klaiman, C., Carter, A. S., & Schultz, R. T. Audiovisual processing in children with and without autism spectrum disorders. Journal autism and dev dis, 2008; 38, 1349-1358.

[7.] Smith, E. G., & Bennetto, L. Audiovisual speech integration and lipreading in autism. J Child Psychol Psychiatry, 2007; 48(8), 813-821.

[8.] Magnée, M. J., De Gelder, B., Van Engeland, H., & Kemner, C. Audiovisual speech integration in pervasive developmental disorder: evidence from event-related potentials. J Child Psychol Psychiatry, 2008; 49(9), 995-1000.

[9.] Rogers, S. J. What are infant siblings teaching us about autism in infancy? Autism Research, 2009; 2(3), 125-137.

[10.] Irwin, J. & DiBlasi, L. (2017). Audiovisual speech perception: A new approach and implications for clinical populations. Lang Linguist Compass. 2017; DOI: 10.1111/lnc3.12237.

[11.] Brancazio, L. Lexical influences in audiovisual speech perception. J Exp Psychol Hum Percept Perform, 2004; 30(3), 445.

[12.] Eigsti, I. M., & Shapiro, T. A systems neuroscience approach to autism: biological, cognitive, and clinical perspectives. Ment Retard Dev Disabil Res Rev, 2003; 9(3), 206-216.

[13.] Alcántara, J. I., Weisblatt, E. J., Moore, B. C., & Bolton, P. F. Speech-in-noise perception in high-functioning individuals with autism or Asperger's syndrome. J Child Psychol Psychiatry 2004; 45(6), 1107-1114, 2004.

[14.] Jerger, S., Damian, M. F., Tye-Murray, N., & Abdi, H. Children use visual speech to compensate for non-intact auditory speech. J. Exp Child Psychol, 2014; 126, 295- 312.

[15.] Irwin, J., Avery, T., Brancazio, L., Turcios, J., Ryherd, K., & Landi, N. Electrophysiological indices of audiovisual speech perception: beyond the McGurk effect and speech in noise. Multisens Res, 2018 *31*(1-2), 39-56.

[16.] Kashino, M. Phonemic restoration: The brain creates missing speech sounds. Acoust Sci Technol, 2006; 27(6), 318-321, 2006.
[17.] Samuel, A. G. Phonemic restoration: insights from a new methodology. J Exp Psychol General, 1981; 110(4), 474.
[18.] Warren, R. M. Perceptual restoration of missing speech sounds. Science, 1970; 167(3917), 392-393.
[19.] C. Lord, M. Rutter, P. C. DiLavore, S. Risi, K. Gotham & S. Bishop, Autism Diagnostic Observation Schedule, Second Edition. Torrance, CA: Western Psychological Services, 2012.
[20.] Rutter, M., Le Couteur, A., & Lord, C. Autism diagnostic interview-revised. *Los Angeles, CA:* Western Psychological Services, 2003 *29*, 30.
[21.] Wechsler, D. WASI-II: Wechsler abbreviated scale of intelligence. Psychological Corporation; 2011.
[22.] van Casteren, M. & Davis, M.H. Match: A program to assist in matching the conditions of factorial experiments. Behavior Research Methods, 2007; 39(4), 973-978.
[23.] Boersma, P. Praat, a system for doing phonetics by computer. Glot international, 2002, *5*.