Auditory-Visual Perception of VCVs Produced by People with Down Syndrome: Preliminary Results

Alexandre Hennequin, Amélie Rochet-Capellan, Marion Dohen

Auditory-visual Perception of VCVs Produced by People with Down Syndrome: Preliminary Results

Alexandre Hennequin, Amélie Rochet-Capellan & Marion Dohen

Univ. Grenoble Alpes, GIPSA-Lab, F-38000 Grenoble, France
CNRS, GIPSA-Lab, F-38000 Grenoble, France
{alexandre.hennequin,amelie.rochet-capellan,marion.dohen}@gipsa-lab.grenoble-inp.fr

Abstract

Down Syndrome (DS) is a genetic disease involving a number of anatomical, physiological and cognitive impairments. More particularly it affects speech production abilities. This results in reduced intelligibility which has however only been evaluated auditorily. Yet, many studies have demonstrated that adding vision to audition helps perception of speech produced by people without impairments especially when it is degraded as is the case in noise. The present study aims at examining whether the visual information improves intelligibility of people with DS. 24 participants without DS were presented with VCV sequences (vowel-consonant-vowel) produced by four adults (2 with DS and 2 without DS). These stimuli were presented in noise in three modalities: auditory, auditory-visual and visual. The results confirm a reduced auditory intelligibility of speakers with DS. They also show that, for the speakers involved in this study, visual intelligibility is equivalent to that of speakers without DS and compensates for the auditory intelligibility loss. An analysis of the perceptual errors shows that most of them involve confusions between consonants. These results put forward the crucial role of multimodality in the improvement of the intelligibility of people with DS.

Index Terms: Down Syndrome, Speech perception, Multimodality, Intelligibility.

1. Introduction

Down Syndrome (DS) is a genetic disease caused by the presence of a third chromosome 21. It is the first genetic cause of intellectual deficiency [1]. Among others, it involves impairments in speech production which cannot be fully accounted for by the intellectual deficiency [2]–[4]. Phenotipic differences observed in people with DS, such as hearing impairments, muscle hypotonia, vocal tract abnormalities etc, also interfer with speech production [4], [5]. All these specificities result in a lower intelligibility. This lack of intelligibility has however only been studied auditorily [2], [3]. On the other hand, a large body of research has shown that seeing someone without DS speak improves the perception of his/her speech, especially when it is deteriorated as is the case in a noisy environment [6], [7]. This study therefore aims at evaluating whether or not adding the visual modality improves the intelligibility of people with DS.

Kent & Vorperian (2013) published a review [4] of the research conducted on the speech of people with DS since the 50s. Despite discrepancies in the results, it appears that fundamental frequency (f0) is generally higher in the speech of people with DS than in that of controls. Vocal quality is also judged as breathy and rough. Acoustically, studies put forward increased perturbations and reduced SNR (Signal-to-Noise Ratio). People with DS tend to produce more articulatory and/or phonological errors [2], [3]. Acoustical analyses of vowel production in vowel-consonant-vowel (VCV) contexts suggest that people with DS display more variability in the first two formants (F1 and F2) and a higher F0 compared with typical speakers ([8]; see also [4]). The production of speech sounds is disturbed by various anatomical specificities [4] : abnormal dentition (number and alignment), high palate, smaller oral cavity resulting in a relatively larger volume occupied by the tongue, etc. Motor control is also affected: generally, people with DS are described as hypotonic. A study on limb motion [9] however suggests that rather than hypotonia, people with DS would rather display higher muscle activation thresholds: initiating movements, and more specifically those required to produce speech, may require more effort for people with DS.The speech of people with DS is therefore altered not only because of intellectual deficiency but by many other factors. This results in an observed lower intelligibility of people with DS, variable across individuals [2], [3].

Intelligibility is defined as "how well a speaker's acoustic signal can be accurately recovered by a listener" [10]. Yet it is well known that the visual information helps perceive speech, and thus improves intelligibility, of typical speakers, especially in adverse conditions such as in noisy environments ([6], [7], [11] ; for a review, see [12]). For example, Grant & Seitz [7] showed that seeing the face of a speaker producing sentences in noise improves the detection threshold of these sentences by 1.6 dB. Summerfield [11] put forward the fact that audition and vision are complementary in speech perception by examining confusions between English consonant in the auditory and visual modalities. It indeed appeared that mode of articulation is the most robust feature in the auditory modality whereas it is place of articulation in the visual modality.

The speech produced by people with DS is acoustically degraded and displays a reduced SNR [4]. This raises the question whether adding the visual information would improve their intelligibility, as is the case for the speech produced by people without DS in noise. Considering the anatomical and motor specificities of speech production in DS, it remains unsure what their visual input consists in and how it is perceived.

In the literature, speech production impairments in DS are often related to dysarthria [3]–[5], [8]. Results in auditory-visual (AV) perception of dysarthric speech could therefore provide valuable insight even if they should not prevail as granted for both populations. Hustad & Cahill [13] examined

the perception of low semantically predictable sentences produced by 5 speakers with dysarthria. Even if overall results show a greater intelligibility in the AV condition than in the auditory only condition, this difference is significant only for the speaker with the most severe dysarthria. A comparable study [14] investigated the perception of speech produced by 8 people with Parkinson's disease and dysarthria. The results show an improvement in intelligibility when vision is added to audition only for 3 out of 8 speakers, i.e. those with the lowest intelligibility scores in the auditory only condition.

The present study aims at analyzing whether adding vision to audition can improve the intelligibility of speech produced by people with DS. The speech of 4 speakers (2 with DS, 2 without DS matched in gender and age) was presented in 3 modalities (Auditory Only, Video Only and Auditory-Visual) to listeners without DS, with little or no experience with this population. Meaningless VCV sequences were used to control for phonetic material and isolate pure segmental intelligibility from semantic context effects.

2. Method

2.1. Participants

24 native speakers of French without DS participated in this study (12 f – age: mean=25.1; se=3). None reported any uncorrected vision problems or any hearing or speech disorders or phonological issue. Before the experiment, their audition was positively tested with a pure-tone hearing screening at 30 dB for 500 Hz, 1 kHz, 2 kHz and 4 kHz bilaterally. At the end of the experiment, they received a 15€ gift card. All had little or no experience with people with DS.

2.2. Speakers and stimuli

4 native speakers were selected from a database recorded in a previous study (see [8]): 2 speakers without DS (a 22 year-old male and a 21 year-old female) and 2 speakers with DS (a 21 year-old male and a 19 year-old female). The speakers were matched in gender and approximate age. 9 expert participants performed an auditory perception pre-test without noise in order to individually evaluate the auditory intelligibility of the speakers with DS. From this test, we chose two speakers with a relatively good intelligibility for the present study: a too poor intelligibility coupled with the fact that stimuli were noised would have led to a floor effect.

The stimuli used consisted of 16 meaningless VCV sequences with V = /a/ and C = {[b], [d], [g], [p], [t], [k], [f], [s], [ʃ], [v], [z], [ʒ], [l], [ʁ], [m], [n]} covering the manners and places of articulation of French. Each VCV sequence was produced three times, and the clearest production (both auditorily and visuallly) was chosen as a stimulus for the present study.

The stimuli were recorded in a soundproof room. Speakers were seated in a chair in front of a loudspeaker, wore a head mounted microphone (Sennheiser HP4) and were filmed using a HD digital camera (Panasonic HC-X920). They were asked to repeat the VCV sequences they heard through the loudspeaker. Audio was sampled at 44100 Hz (Focusrite Scarlett 6i6 soundcard). Each audio file was normalized at 70 dB using Praat and noised (signal to noise ratio = -4 dB) with a "cocktail party" noise (BDBRUIT, [15]). The files were then resynchronized with the videos using FFmpeg (https://www.ffmpeg.org/) at a 960x540 pixels resolution in three different versions: Auditory only (A, with the picture of a loudspeaker), Video only (V) and Auditory-Visual (AV) resulting in a total of 192 stimuli: 4 speakers x 3 modalities x 16 VCV sequences.

2.3. Procedure

Participants were seated approximately at 60 cm from a 24" screen and wore a headset (Audio Technica BPHS1). The experience was programmed using the Psychophysics Toolbox [16]–[18]. Videos were presented at the center of the screen. The test was divided into three blocks, one for each modality (A, V and AV), consisting of 64 stimuli each (16 VCV x 4 speakers). Orders of blocks and stimuli within each block were randomized. Participants were informed that they would either hear/ or see or hear and see a stimulus presented twice in a row. Their task was then to repeat what they had understood. They then hit a key on the keyboard to move to the next stimulus. Audio was sampled at 48 kHz (Focusrite Scarlett 6i6 soundcard). They were told that the stimuli were meaningless speech sequences. No other information on the structure of the sequences was provided. The test was preceded by a training phase using noiseless stimuli different from those of the experiment. Training consisted of 2 stimuli per block. The order of presentation of the blocks was not necessarily the same in the training and the actual experience. At the end of this training, participants listened to a brief sample of the "cocktail party" noise and were informed that the stimuli from the actual experiment were all mixed with this type of noise.

2.4. Response transcription and analyses

The responses provided by the participants were transcribed using the following code: BeforeV1-V1-C-V2-AfterV2. Each part was phonetically transcribed or left empty (e.g., "brata" instead of "ata": beforeV1='br' – V1='a' – C='t' – V2='a' – AfterV2=''). An unperceived consonant was transcribed 'h'. If the response was only a vowel (e.g., 'a' instead of 'ata'), it was transcribed as V2 (V1='' – C = 'h' – V2='a'). Responses that could not be transcribed were coded '?' (BeforeV1=V1=C=V2=AfterV2='?').

Results were analyzed using the R software (R Development Core Team, 2008) and analyses of variance (aov function with default parameters). Post-Hoc comparisons (Student tests) were corrected using the Bonferroni correction.

Confusion trees were built to examine confusions between consonants in the visual and auditory modalities separately. First confusion matrices were computed for each speaker group : each cell (m_i, n_j) contains the number of times the consonant m_i was perceived as n_j. These matrices were then used to compute Euclidian distance matrices which were in turn fed to a divisive hierarchical clustering algorithm (diana function in the package cluster from R - [19]). The resulting data was then plotted as a dendogram.

3. Results

A correct response corresponds to the case in which V1, C and V2 were correctly identified and BeforeV1='' and AfterV2=''. Global analyses show that 44.2% of the responses were correct, and 54.4% contained at least one error. The remaining 1.4% could not be transcribed. Speaker group and modality do not have an effect on these percentages ($p > 0.1$).

3.1. Correct responses

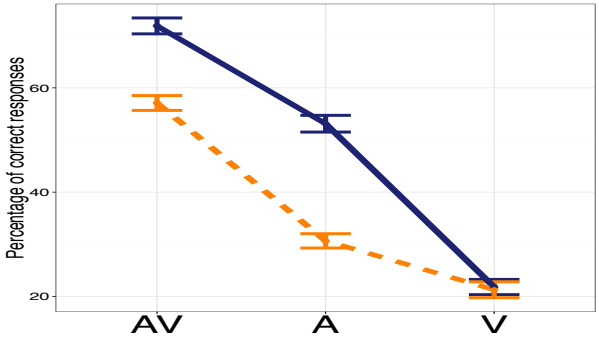

Figure 1: Percentages of correct responses depending on speaker group (without vs. with DS) and modality (AV, A, V)

Figure 1 shows the percentages of correct responses depending on speaker group and modality. We conducted an ANOVA on these percentages with two within-subject factors (speaker group and modality) and one between-subject factor (order of presentation of modalities). Modality has a significant effect ($F(2,36)=263.5$ – $p < 0.001$) : there are more correct responses in the AV modality than in A (A vs. AV: $t(23)=-12.1$; $p < 0.001$) than in V (A vs. V: $t(23)=5.5$ $p < 0.001$). Overall, results are better for speakers without DS than speakers with DS ($F(1,18)=14.6$ – $p = 0.001$). This however depends on the modality (speaker group * modality: $F(2,36) = 13.6$ – $p < 0.001$) : in the A modality, speakers without DS are better perceived than those with DS ($t(23) = 6.7$ – $p < 0.001$) but this is not the case in the V modality ($t(23) = -0.9$ – $p > 0.9$). In the AV modality, there is only a trend towards significance (significant before correction) for speakers without DS to be better perceived than those with DS ($t(23)=2,1$ – $p > 0.1$). The order of presentation also has a significant effect ($F(5,18)=3,2$ – $p < 0.05$) and interacts with modality ($F(10,36)=3,2$ – $p < 0.01$). This effect is significant only in the V modality: when the AV modality was presented before V, the results in V are significantly better.

3.2. Errors

General analysis – Errors were divided into three categories: insertions before and/or after the VCV sequence (beforeV1≠'' and/or afterV2≠''), errors on vowels (V1 and/or V2) and error on the consonant. These three categories can co-occur in the same response. Figure 2 shows the repartition of the errors depending on error type.

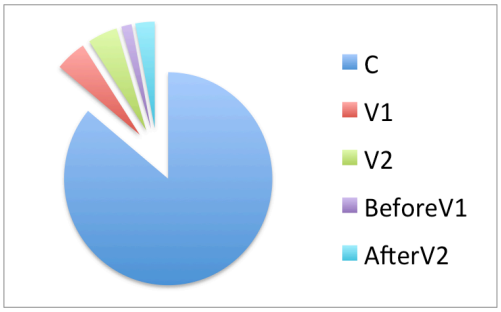

Figure 2: Distribution of errors depending on location.

More than 80% of the errors (response contains one or more errors) occur on the consonant (Figure 2). This is the reason why only errors on the consonant were analyzed into detail.

Errors on the consonant – The consonant (C) responses were divided into three categories: correct, confusion with another consonant and other (e.g., addition of one or more consonants). An ANOVA was conducted on this data with three within-subject factors: speaker group, modality and error type. Note that 94.5% of the responses are either correct responses or confusions with another consonant.

Confusion with another consonant is the most frequent type of error ($F(1,23) = 228.3$ – $p < 0.001$). Errors on the consonant occur more often in A and V modalities than in AV ($F(2,46) = 206.4$ – $p < 0.001$). Overall, there is no difference between speaker groups ($F(1,23) = 2.4$ – $p > 0.1$). There are however more errors in the A modality for speakers with DS than for speakers without DS (speaker group * modality: $F(2,46) = 3.6$ – $p < 0.05$).

In order to further analyze confusions between consonants, we used the same type of representation as Summerfield [11]. Confusions are presented as a tree in which each "leaf" corresponds to a consonant. The closer the consonants are in the tree, the more often they are confused. Figure 3 presents the confusion trees depending on modality (A vs. V) and speaker group (without vs. with DS). The colors used in the trees for the V modality correspond to a grouping of consonants by place of articulation (1 color for each place of articulation) and those in the tress for the A modality to a grouping by manner of articulation (voiced vs. unvoiced). Figure 3, shows that consonants are grouped by color (manner of articulation) for speakers without DS but not for those with DS. Manner of articulation is thus a robust feature in the A modality for speakers without DS but not for those with DS. In the V modality, consonants are mainly grouped by place of articulation (consonants with the same color are grouped) in both speaker groups. The only exception is for velars for speakers without DS and for velars and alveolars for speakers with DS. The place of articulation therefore appears to be a robust feature in the V modality for both groups of speakers.

4. Discussion

This study examined the perception of VCV sequences produced by speakers with and without DS by participants without DS in three modalities: auditory (A), visual (V) and auditory-visual (AV). The sequences were presented in a cocktail-party noise. The aim of this study was to question whether adding the V modality could improve the intelligibility of speech produced by speakers with DS as it does for speech produced by speakers without DS.

Overall, the results show that adding the visual modality improves the perception of noisy speech (better results in the AV modality than in the A and V modalities) whichever the group. The percentage of correct responses are lower for speakers with DS than for those without DS in the A modality. This confirms previous results ([2], [3]): in the A modality, speech produced by our two speakers with DS is less intelligible than that produced by speakers without DS. The difference in intelligibility between the two groups of speakers is much smaller in the AV modality (not even significant after correction for multiple comparison). This suggests that adding the visual modality at least partly compensates for the lack of auditory intelligibility of speech produced by our two speakers with DS. There is no difference in the percentages of correct responses between the two groups in the V modality: our two speakers with DS are visually as intelligible as those without DS. Note however that when the V modality is presented

before the AV modality in the experimental procedure, the results are significantly lower than when it is presented after the AV modality for both speaker groups. This presentation order effect could be explained by the fact that, since it is easier to perceive speech in the AV modality, doing it before perceiving it in the V modality trains the participants and improves their perception in the V modality. It is also possible that participants memorize the auditory-visual association of a given stimulus, and that, when they encounter it later in the experiment in V modality, it is easier for them to recover the correct response.

For dysarthric speech, an improvement in intelligibility by adding the V modality was observed only for speakers with the most severe dysarthria [13], [14]. In this study we found such an improvement for our two speakers with DS with a relatively good level of intelligibility (see 2.2 for more information). It would therefore be interesting to explore whether the improvement is even larger for speakers who are less intelligible.

Over all speaker groups, the most frequent type of error was confusion between consonants. Note that the consonant was the only phoneme of the sequence that varied in the experiment (V = /a/). Further analysis of the types of confusions made in the A and V modalities (confusion trees) confirmed that, for speakers without DS, the most robust feature in noise is the manner of articulation in the A modality and the place of articulation in the V modality (same result as in [11]). Our results show that the robustness of place of articulation in V perception is relatively well preserved in the speech produced by people with DS. Robustness of manner of articulation in A perception is however altered in the speech of people with DS. This suggests that voicing is a feature especially difficult to produce by people with DS as already suggested in [8].

More detailed analyses will have to be conducted to examine the ability to detect place and manner of articulation in the speech of people with DS compared to that of people without DS. Note that this study used only one vowel. We indeed decided to focus on the perception of consonants in four speakers and adding a variation of the vowel would have resulted in a too long experiment. It would however be interesting to examine whether adding the V modality also improves the perception of vowels produced by people with DS especially since we know that they are acoustically different from those produced by speakers without DS [8].

5. Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no.339152- "Speech Unit(e)s"). This work is linked to the project "Communiquons Ensemble" (Communicate Together) funded by the Fondation Internationale de la Recherche Appliquée sur le Handicap (FIRAH). The authors want to thank the participants, the Association de Recherche et d'Insertion Sociale des Trisomiques (ARIST) and the ARIST professionals of the ESAT-SAJ.

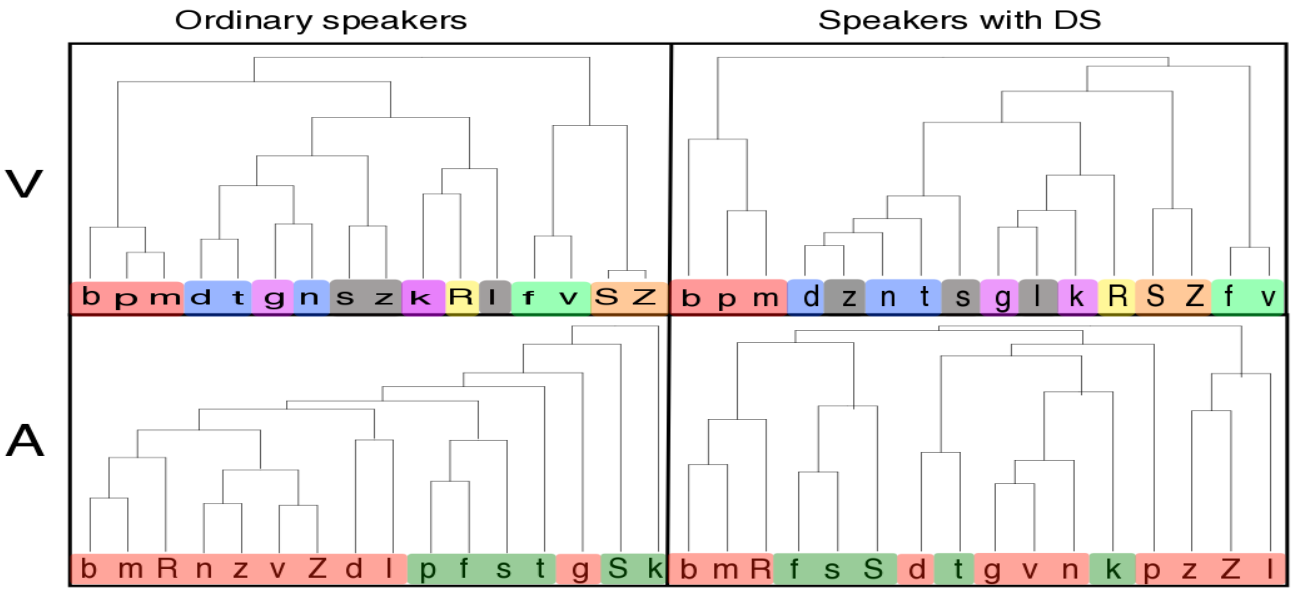

Figure 3: Consonant confusion trees depending on speaker group (without DS and with DS) and modality (Visual and Auditory). The colors in the trees for the V modality correspond to grouping of consonants by place of articulation (red=bilabial – green=labiodental – bleu=dental – grey=alveolar – orange=post-alveolar – pink=velar – yellow=uvular) and those in the trees for the A modality to a grouping by manner of articulation (red = voiced – green=unvoiced). N.B.: S=[ʃ] – Z=[ʒ] – R=[ʁ].

References

[1] G. Katz and E. Lazcano-Ponce, “Intellectual disability: definition, etiological factors, classification, diagnosis, treatment and prognosis,” *Salud Publica Mex*, vol. 50 Suppl 2, pp. s132–41, 2008.

[2] L. Kumin, “Speech intelligibility and childhood verbal apraxia in children with Down syndrome.,” *Downs. Syndr. Res. Pract.*, vol. 10, no. 1, pp. 10–22, 2006.

[3] K. Bunton, M. Leddy, and J. Miller, “Phonetic intelligibility testing in adults with Down syndrome,” *Down Syndr. Res. Pract.*, vol. 12, no. 1, pp. 1–4, 2007.

[4] R. D. Kent, H. K. Vorperian, J. Kreiman, and B. A. M. Maassen, “Speech Impairment in Down Syndrome: A Review,” *J. Speech, Lang. Hear. Res.*, vol. 56, no. 1, pp. 178–210, 2013.

[5] G. E. Martin, J. Klusek, B. Estigarribia, and J. E. Roberts, “Language Characteristics of Individuals with Down Syndrome,” *Top. Lang. Disord.*, vol. 29, no. 2, pp. 112–132, 2009.

[6] Q. A. Summerfield, “Use of Visual Information for Phonetic Perception,” *Phonetica*, vol. 36, no. 4–5, pp. 314–331, 1979.

[7] K. W. Grant and P. F. Seitz, “The use of visible speech cues for improving auditory detection of spoken sentences.,” *J. Acoust. Soc. Am.*, vol. 108, no. 3 Pt 1, pp. 1197–1208, 2000.

[8] A. Rochet-capellan and M. Dohen, “Acoustic characterisation of vowel production by young adults with Down syndrome,” *ICPhS*, 2015.

[9] M.Latash, L. Wood, and D. Ulrich, “What is currently known about hypotonia, motor skill development, and physical activity in Down syndrome,” *Down Syndrome Research and Practice (Online)*. 2008.

[10] K. C. Hustad, “The relationship between listener comprehension and intelligibility scores for speakers with dysarthria,” *JSLHR*, vol. 51, no. 3, pp. 562–573, 2008.

[11] Q. Summerfield, “Some preliminaries to a comprehensive account of audio-visual speech perception,” in *Hearing by Eye: The Psychology of Lip-reading*, 1987, pp. 3–51.

[12] M. Dohen, “Speech through the ear, the eye, the mouth and the hand,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5398 LNAI, pp. 24–39, 2009.

[13] K. C. Hustad and M. A. Cahill, “Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech,” *American Journal of Speech-Language Pathology*, vol. 12. pp. 198–208, 2003.

[14] C. K. Keintz, K. Bunton, and J. D. Hoit, “Influence of visual information on the intelligibility of dysarthric speech.,” *Am. J. Speech. Lang. Pathol.*, vol. 16, no. 3, pp. 222–34, 2007.

[15] J. Zeiliger, J. Serignat, D. Autessere, and C. Meunier, “Bd_bruit, une base de données de parole de locuteurs soumis à du bruit,” *Actes des Xèmes JEP*, pp. 287–290, 1994.

[16] D. H. Brainard, “The Psychophysics Toolbox.,” *Spat. Vis.*, vol. 10, no. 4, pp. 433–436, 1997.

[17] D. G. Pelli, “The VideoToolbox software for visual psychophysics: transforming numbers into movies.,” *Spatial Vision*, vol. 10, no. 4. pp. 437–442, 1997.

[18] M. Kleiner, D. H. Brainard, D. G. Pelli, C. Broussard, T. Wolf, and D. Niehorster, “What’s new in Psychtoolbox-3?,” *Perception*, vol. 36, p. S14, 2007.