Cue integration and context effects in speech: Evidence against speaking rate normalization

Supplemental Material

Joseph C. Toscano[1,2] and Bob McMurray[2,3]

[1]Beckman Institute, University of Illinois at Urbana-Champaign
[2]Delta Center, University of Iowa
[3]Dept. of Psychology and Dept. of Communication Sciences & Disorders, University of Iowa

The difference in the size of effects between synthetic and natural speech could be due to the presence of additional cues in the natural stimuli that covary with the primary continuum dimension, VOT (Shinn, Blumstein, & Jongman, 1985), and cue-integration principles may be able to account for these differences. Here, we examine this possibility by first conducting a phonetic analysis of our stimuli to determine the extent to which cues covary and then running simulations with the weighted Gaussian mixture model (WGMM; Toscano & McMurray, 2010) to examine their effect.

S1. Acoustic measurements

One additional cue that may have been present in the naturally-produced stimuli is F1 frequency at the onset of voicing. Listeners have been shown to use F1 onset as a cue to voicing (Summerfield & Haggard, 1977), and in the naturally-produced stimuli used in Experiment 2, the voiced tokens has a mean F1 onset of 413 Hz (SD=137 Hz across words), while the voiceless recordings had an average of 673 Hz (SD=282 Hz across words). Since our stimuli were created by cutting back the amount of voicing, F1 onset frequency may have been correlated with VOT. Figure S1 shows a schematic of the procedure used to create the natural VOT continua.

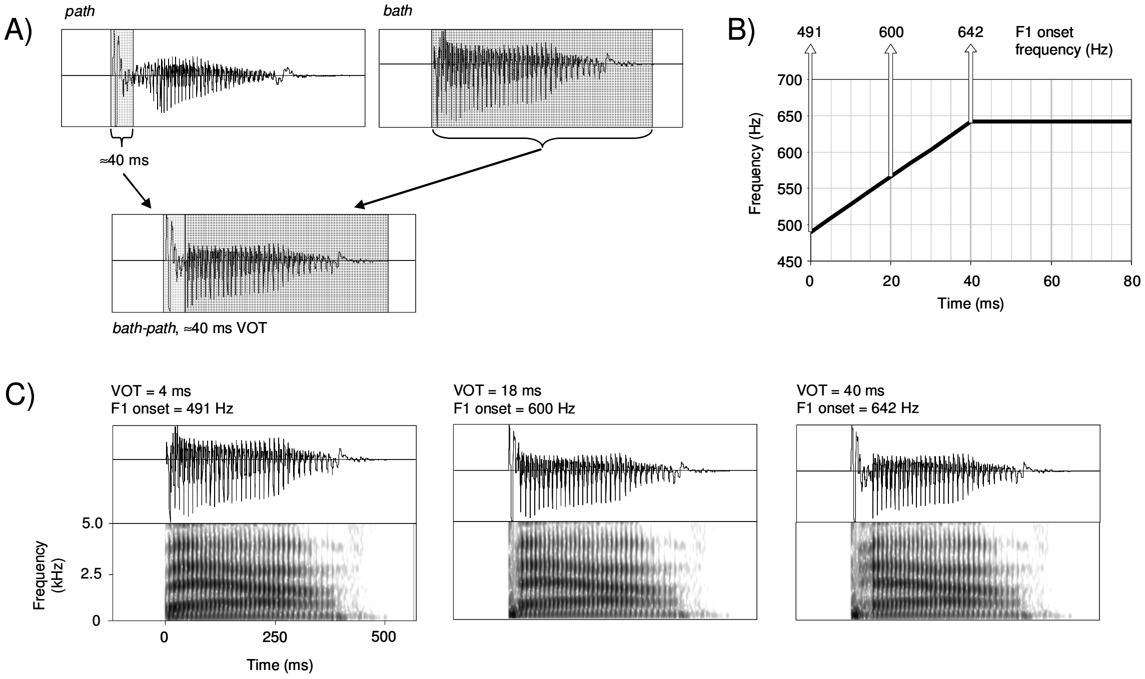

Figure S1. (A) Schematic representation of VOT cutback procedure for creating naturally-produced stimuli. Portions of aspiration from the voiceless token are spliced into the voiced token in several millisecond steps, while a corresponding duration is removed from the onset of the voiced token, creating a series of VOTs. This procedure causes formant frequencies at the onset of voicing to co-vary with VOT (if there is any formant transition). (B) Schematic of the first formant of *bath* with cut points for steps 1, 5, and 9, and the corresponding F1 onset frequencies for those steps. (C) Waveforms and spectrograms for steps 1, 5, and 9 from the *bath-path* continuum.

F1 generally rises for bilabials, from the release to its value at the vowel centroid. Since VOT was manipulated by cutting increasingly larger segments from the onset of the voiced token, the F1 frequency at that cut point (i.e., the onset of voicing) will increase as VOT increases. Since higher F1 values lead to more voiceless responses (as do longer VOTs), VOT and F1 at onset may be correlated, creating stimuli not unlike those of Shinn et al. (1985). Listeners may use F1 at onset in the natural stimuli in addition to VOT and VL, reducing the apparent size of the VL effect. A similar pattern is likely to be seen in F2 and F3, as these also rise in bilabials (though there is less evidence that they are involved in voicing judgments). The synthetic stimuli in Experiment 1 were created using a similar approach (cutting back the onset of AV) and also contained rising formant frequencies at onset. However, these stimuli used steep formant transitions, and as a result, cutting back voicing may not have introduced substantial changes in formant onset frequency. Thus, these cues may bias listeners' responses less in synthetic speech.

To test this, we measured formant onset frequencies for the continua from Experiments 1 and 2. For each stimulus item, the frequency of the first three formants was measured at the onset of voicing. The stimuli were measured using the formant tracker in Praat (Boersma & Weenink,

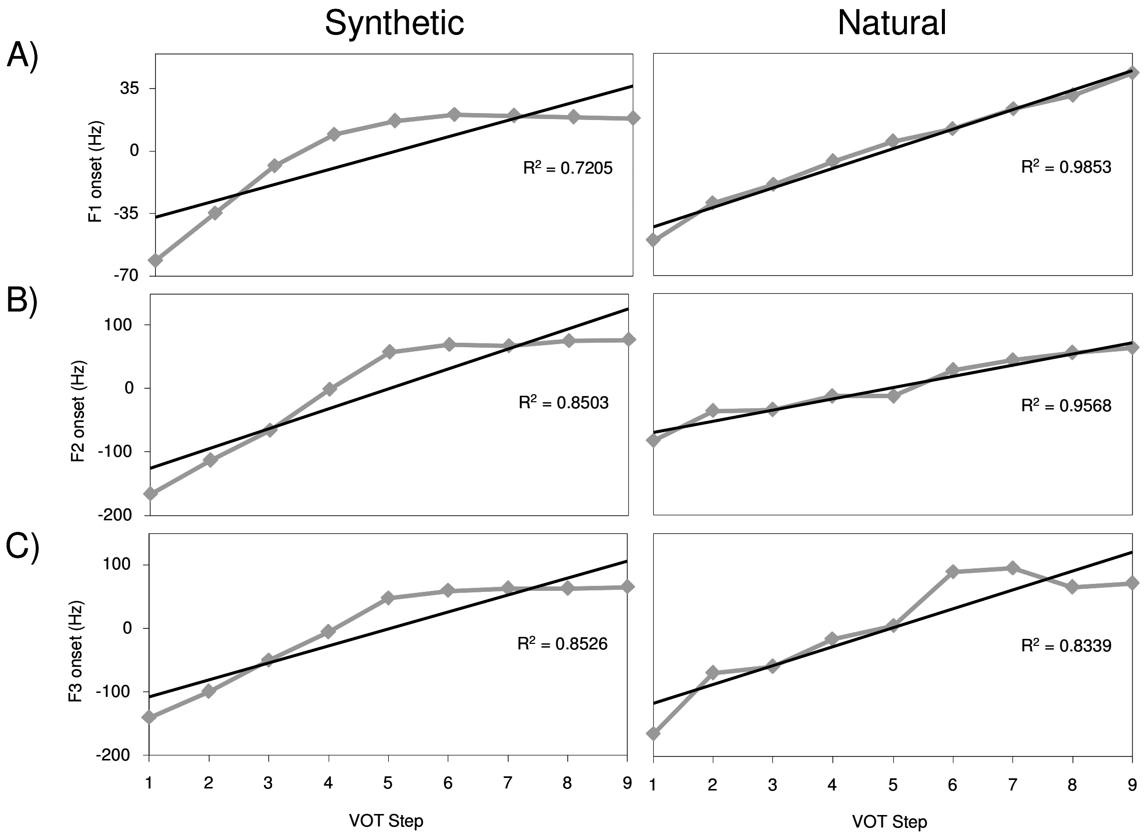

Figure S2. Formant onset frequencies for F1 (panel A), F2 (panel B), and F3 (panel C) as a function of VOT step and stimulus type (left side: synthetic; right side: naturally-produced). Formant onset values were centered so that they had a mean of zero for each continuum and were then averaged. For natural stimuli, the onsets of the first two formants were correlated with VOT across the continuum, and F3 showed a linear effect for steps 1-6. For synthetic stimuli, formant onsets increase with VOT for steps 1-4, but are flat for steps 5-9.

2009) with the Split-Levinson algorithm (Willems, 1986). For each word, the formant onset was measured as the mean formant value in a 15-ms time window starting at the visible onset of voicing in the spectrogram. A custom MATLAB program was then used to visually inspect the formant tracks computed by Praat and adjust them if they did not match the visible formants in the spectrogram. Values from the two long and short VL stimuli were averaged together, though there were no major differences between the two for a given word and VOT.

Figure S2 shows the mean onset frequencies of the first three formants as a function of VOT. To cope with the three missing stimuli (see Note 5 in the paper), we centered the data by subtracting the mean formant frequencies from the formant frequencies at each step. This was done separately for each continuum and stimulus type (natural and synthetic).

As expected, formant onsets co-varied with VOT more in the naturally-produced continua than in the synthetic continua. For F1 and F2, naturally-produced stimuli show a linear effect across the VOT continuum, and F3 showed a linear effect from steps 1 to 6; it remained mostly constant from steps 7 to 9. In contrast, the synthetic stimuli were flat from about step 5 to step 9. Additionally, the difference in F1 between the VOT endpoints was 94 Hz for naturally-produced continua and 80 Hz for synthetic continua. Thus, not only did the naturally-produced continua have stronger covarying cues, but they also showed a larger magnitude of frequency change. These results fit with the prediction that they contained more informative formant onset cues than the synthetic stimuli.

These acoustic measurements help explain why we observed a smaller effect of VL in Experiment 2: stimuli containing cues that covary with VOT produce smaller VL effects. This implies that listeners might reweight VL as a function of which other cues are available. However, in many cue integration models (Oden & Massaro, 1978; Nearey, 1997; Toscano & McMurray, 2010) the effects of multiple cues are additive such that if two cues covary, this can increase the apparent effect of one and reduce the apparent effect of cues that are not-covarying, even without changing cue weights. To examine this possibility, we simulated the effects of additional cues using the WGMM.

S2. Simulations

Given the acoustic analyses, we now ask whether similar differences between stimuli produce the same results in a computational model of cue integration, specifically, the weighted Gaussian mixture model (WGMM; Toscano & McMurray, 2010). In this model, categories are represented as Gaussian distributions defined by their mean (prototype values) and variance (their spread in cue-space). Prior work with this type of model (McMurray, Aslin, & Toscano 2009; de Boer & Kuhl, 2003; Vallabha et al., 2007; Lake, Vallabha, & McClelland, 2009) has generally focused on its ability to learn categories along specific acoustic-phonetic dimensions. Here, we use it as a platform in which to understand multiple cue integration.

While it is possible to model categories defined by multiple cues using multidimensional Gaussians (e.g., Vallabha et al., 2007), this does not allow for an easy examination of cue weights as they are not instantiated explicitly in the model, and it is not necessarily needed to model multiple cues to the same phonological feature (e.g., VOT and VL as cues to voicing). However, Toscano and McMurray (2010) recently extended the Gaussian mixture model approach to weight and combine cues based on their reliability. In this weighted GMM, individual cues are weighted and summed to yield a composite dimension (e.g. VOT and VL are combined to form a voicing dimension) on which categories based on information from multiple cues can be learned. Each individual cue is weighted by its ability to support a given phonological distinction by taking into account the variability within and between categories along that dimension, as well as the likelihood of the categories.

A detailed description of the model architecture is given in Toscano & McMurray (2010). In brief, the model consists of two levels. First, at the cue-level, Gaussians corresponding to each phonological category are estimated along each individual cue dimension (e.g., VOT and VL). These are used to compute the weight of each cue. The individual cue-values are then centered along the dimension, multiplied by these weights, and summed to create a new dimension (e.g. voicing) reflecting the weighted contributions of all the cues. Finally, a new set of Gaussians corresponding to phonological categories are estimated along this combined dimension.

The model's categorization responses (computed from this combined mixture) correspond to those produced by human listeners. Crucially, the model can be tested with different stimuli while holding cue weights constant, allowing us to determine if differences in the effect of VL can be observed without changes in cue weights. Here, we present simulations in which the model is trained on three cues to voicing: VOT, VL, and F1 onset frequency. The VOT/VL relationship for the model was measured with a simulated experiment in which, after training, F1 onset either covaried with VOT (as in the naturally-produced stimuli) or was held constant at an ambiguous value (as in the synthetic stimuli).

Our goal here is not a complete presentation of the model, which can be found in Toscano and McMurray (2010). Rather, we use the model to understand how covarying cues like F1 can contribute to changes in the VL effect.

Architecture

The model consists of Gaussian mixtures for each cue dimension that are used solely to combine and weight cues into a continuous factor (e.g., voicing as a continuous property of the input). This combined factor serves as input to another mixture corresponding to a more abstract phonological dimension. In both types of mixtures, a category (i) is defined by a Gaussian distribution with three parameters: its probability of occurrence (ϕ), mean (μ), and standard deviation (σ). The likelihood of a particular cue-value (x) for a given category in the mixture is

$$G_i(x) = \phi_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right) \tag{1}$$

The overall likelihood for that cue-value is the sum of the likelihoods from each category,

$$M_m(x) = \sum_i^K G_i(x) \tag{2}$$

where K is the number of Gaussians in the mixture.

The model is given cue-values as input to each cue mixture. These inputs are normalized by converting them to z-scores using a measure of the mixture's central tendency (*c*),

$$c = \frac{\sum_i^K \mu_i \phi_i}{\sum_i^K \phi_i} \tag{3}$$

and variability (*v*),

$$v = \sqrt{\frac{\sum_i^K \left((\mu_i - c)^2 + \phi_i \sigma_i^2 \right)}{\sum_i^K \phi_i}} \tag{4}$$

Normalized inputs (*z*) are computed according to

$$z = \frac{x - c}{v} \tag{5}$$

and cue dimension weights (*w*) are computed according to

$$w = \left(\sum_i^K \sum_j^K \frac{\phi_i \phi_j (\mu_i - \mu_j)^2}{\sigma_i \sigma_j} \right) / 2 \tag{6}$$

Weights are normalized so that they sum to one, and the input for the phonological mixture is computed by summing the weighted, normalized cue inputs according to

$$x_p = \sum_c^D w_c z_c \tag{7}$$

where *D* is the number of cue dimensions. Parameters in the model are updated to reflect the distributional statistics of the input using maximum likelihood estimation via gradient descent. The learning rules for each parameter in a single mixture are:

$$\Delta \phi_i = \eta_\phi \frac{G_i(x)}{M(x)} \tag{8}$$

$$\Delta \mu_i = \eta_\mu \left(\frac{G_i(x)}{M(x)} \right) \frac{(x - \mu_i)}{\sigma_i^2} \tag{9}$$

$$\Delta \sigma_i = \eta_\sigma \left(\frac{G_i(x)}{M(x)} \right) \left(\sigma_i^{-3} (x - \mu_i)^2 - \sigma_i^{-1} \right) \tag{10}$$

where η is the learning rate for each parameter. ϕ-values are normalized so that they sum to one, and winner-take-all competition is used to only update the ϕ parameter of one Gaussian on each training trial. This allows the model to eliminate unnecessary categories in the mixture (by decreasing their ϕ-values to near zero) to determine the number of categories in the input.

For the simulations presented here, the model had three cue-level mixtures (VOT, VL, and F1 onset) and one phonological-level corresponding to the voicing dimension.

Training data and model parameters

The model was trained on cue-values randomly sampled from distributions based on two large sets of phonetic measurements (the VOT distributions given by Lisker and Abramson, 1964; and the VL distributions given by Allen and Miller, 1999), as well as a distribution of F1 onset values based on the acoustic measurements presented above (voiced mean: 260 Hz, voiceless mean: 300 Hz, SD for both categories: 10 Hz). The sources of the training data were chosen based on the phonetic measurements that were available to us at the time.

The cue-level mixtures were initialized by setting ϕ-values to be equal; setting σ-values to 3, 10, and 3 for the VOT, VL, and F1 mixtures, respectively; and by setting μ to random values sampled from distributions with widths of 75 and means of 25, 179, and 280, for the VOT, VL, and F1 mixtures. For the combined mixture, initial ϕ-values were equal, σ-values were set to 0.03, and μ-values were randomized based on a distribution with a mean of 0 and a width of 0.5.

The number of Gaussians (*K*) in each mixture was 15 (more than it would ultimately need), and these categories were pruned over training as the likelihoods of the unnecessary ones approached zero (see McMurray et al., 2009, for a discussion of this process). Learning rates for the cue mixtures were set to 0.001, 1, and 1 for η_{ϕ}, η_{μ}, and η_{σ}, respectively. Learning rates for the combined mixture were 5^{-5}, 1^{-5}, and 1^{-5}. Each model was trained on 90,000 data points before testing, and 50 repetitions were run.

The number of categories in the model at the end of training was measured by finding the number of Gaussians in each mixture with a likelihood greater than 0.1. Models that over-generalized and had only a single category in any of the mixtures, or that had no above-threshold categories in the combined mixture at the end of training were not included in the analyses.

Testing

The model was tested in two conditions: a covarying F1 condition in which F1 values during testing covaried with VOT, and a constant F1 condition in which they were all set to ambiguous values. In both conditions, models were tested on a VOT continuum ranging from 0 to 40 ms in 5-ms steps and VLs of 125 and 225 ms, similar to those used in Experiments 1 and 2. In the covarying-F1 condition, F1 onset frequencies varied with VOT from 240 to 320 ms in 10-ms steps. In the constant-F1 condition, F1 onsets were held constant at 280 ms, an ambiguous value at the boundary of the two voicing categories defined by the F1 onset values. Cue weights and other parameters were held constant across these two test conditions.

The model's categorization responses for a given set of cues were computed by measuring the posterior probability for each category in the combined mixture and calculating Luce choice ratios (Luce, 1959; temperature = 1) for the proportion of /p/ responses.

Results

Overall, the models were successful in learning voicing categories from the cues. Of the 50 repetitions that were run, two were excluded because they over-generalized and two were excluded because they had no above-threshold categories in the combined mixture. The model captured the distributional statistics of the input quite accurately for the VOT and F1 onset cues. For the VL dimension, the categories that the model learned were close to those in the dataset, but were further apart. This exaggerated the effect of VL (see Toscano & McMurray, 2010, for further discussion.). Figure S3 shows the models' categorization responses for each F1 onset condition. The size of the VL effect is larger for the constant-F1 condition than the covarying-F1 condition. This mirrors the differences in listeners' categorization responses across Experiments 1 and 2, and it fits with the prediction that additional covarying cues in natural speech can reduce the size of the VL effect.

Discussion

The results of the simulation demonstrate that a cue-integration model can account for the differences in the VL effect between natural and synthetic speech without changing the weights of the cues. While this model had explicit weights for each cue, we did not need to manipulate these weights to observe differences in the VL effect due to the acoustic differences between natural and synthetic speech. That is, the model does not need to detect that it is hearing more synthetic-sounding speech and increase the weight of VL. In the constant-F1 condition, only VOT and VL consistently vary with voicing, and formant onsets are not informative (they are at ambiguous values). Thus, categorization responses are dictated by VOT and VL. In contrast, in the covarying-F1 condition, formant onset provides an informative cue to voicing that covaries with VOT. This effectively makes variation along the VOT dimension *appear* to be stronger because it corresponds to variation in multiple cues (VOT and formant onsets) rather than just one. Because the final response is a weighted sum of all available information and an additional cue covaries with VOT, the relative contribution of VL is reduced.

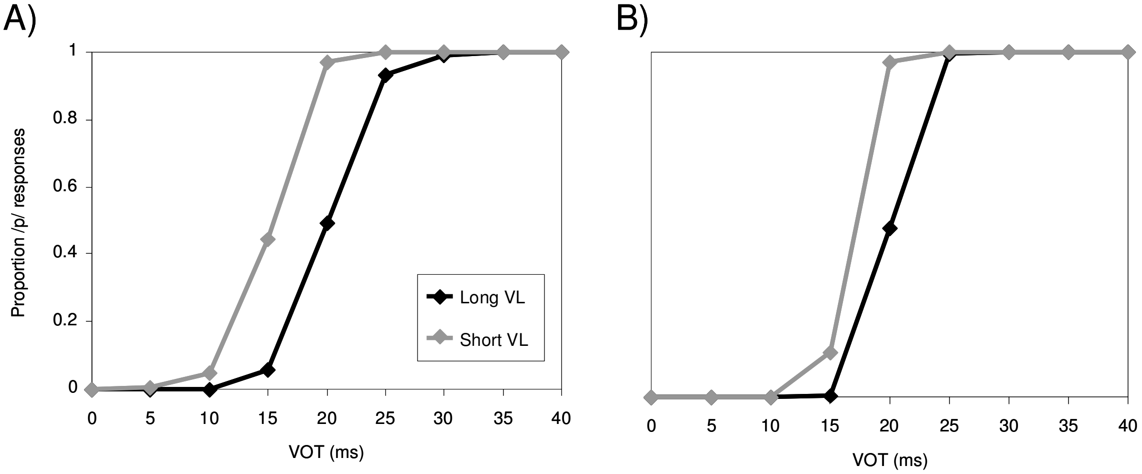

Figure S3. (A) Model categorization responses when additional cues are held constant at an ambiguous value. (B) Model responses when those cues covary with VOT.

References

Allen, J. S., & Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *Journal of the Acoustical Society of America, 106,* 2031-2039.

Boersma, P., & Weenink, D. (2009). Praat: doing phonetics by computer. Available at : http://www.praat.org.

de Boer, B., & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustic Research Letters Online, 4*, 129-134.

Lake, B. M., Vallabha, G. K., & McClelland, J. L. (2009). Modeling unsupervised perceptual category learning. *IEEE Transactions on Autonomous Mental Development, 1*, 35-43.

Lisker, L., & Abramson, A. S. (1964). A cross-linguistic of voicing in initial stops: Acoustical measurements. *Word, 20,* 384-422.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science, 12*, 369-378.

Nearey, T. M. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America, 101*, 3241-3254.

Oden, G. C., & Massaro, D. W. (1978). Integration of feature information in speech perception. *Psychological Review, 85*, 172-191.

Shinn, P. C., Blumstein, S. E., & Jongman, A. (1985). Limitations of context conditioned effects in the perception of [b] and [w]. *Perception & Psychophysics, 38,* 397-407.

Summerfield, Q., & Haggard, M. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America, 62*, 435-448.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science, 34,* 436-464.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences, 104*, 13273-13278.

Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition, 103*, 147-162.

Willems, L. (1986). Robust formant analysis. *IPO Report, 529*, 1-25. Eindhoven: Institute for Perception Research.