Asymmetries in the perception of synthesized speech

Anna C. Janska[1], *Erich Schröger*[2], *Thomas Jacobsen*[3], *Robert A. J. Clark*[4]

[1]IMPRS NeuroCom, University of Leipzig, Germany
[2] University of Leipzig, Germany
[3] Helmut Schmidt University, Hamburg, Germany
[4]CSTR, The University of Edinburgh, U.K.

janska@rz.uni-leipzig.de, robert@cstr.ed.ac.uk

Abstract

It was previously observed [1] that the order of presentation of paired stimuli influenced the number of *different* responses in same-different tasks in speech synthesis evaluation. This paper investigates this phenomenon within the context of cognitive psychology and demonstrates that, as the cognitive psychology literature suggests, there is an effect relating to the prototypicality of the stimulus.

Index Terms: speech synthesis, evaluation, perception, Blizzard Challenge

1. Introduction

When listeners are asked to decide whether two paired stimuli sound equally natural, they are basically asked to position the stimuli in their mental space in relation to their abstract representation of *naturalness* [2]. Then these two distances are gauged and compared, to decide whether two stimuli are equally good samples of the category *naturalness*. This task is not trivial and subject to perceptual biases: Rosch [3] introduced the idea that categories are graded, i.e. that some category members are more representative of a category than others, and that discriminative ability is influenced by a stimulus' distance to the prototype or reference level[1]. Furthermore, she found that stimuli close to the reference point are assimiliated to the reference stimulus to a higher degree than stimuli further away from it [4].

Such warping of the perceptual space has been observed in a systematic manner in different modalities as well as stimulus types: using **linguistic hedges**, Rosch [4] found that given a sentence such as "A ____ is almost a ____", participants would insert the less prototypical item in the first, and the prototypical item in the second slot. This effect was found reliably across 6 stimulus sets. From this order effect Rosch deduced that there was an asymmetrical relation between reference stimuli and somewhat deviant stimuli. These order effects were absent in neutral stimuli. In **harmonies**, Schellenberg found that shifts from in-tune sequences to out-of-tune sequences were more noticable than vice versa [5]. In **tone durations**, Hellström found that repeated presentation of a reference tone could establish that tone as a reference level; deviations from the reference tone length were detected more easily if they followed the reference tone, rather than preceded it. Hellström argues that order effects are more than a bias, and not simply additive [6].

If this order effect also influences the evaluation of synthesized speech, this has some strong implications when employing pairwise comparisons in evaluation studies, such as multidimensional scaling: It is vital that in evaluations, the order of comparisons is carefully balanced, as neglect of this can create a strong bias in the data. This re-emphasizes the need for concise and exhaustive testing and justifies the additional cost of time and money this requires. Furthermore should findings of such an asymmetry rekindle discussions about whether and if so how aggregation across listeners as well as stimuli can be done to generate replicable and valid results. It also raises the question: is any order effect small enough that averaging across both orders is a sufficient way to deal with any effect? And whether anything can be learned from the order effect itself?

2. Modelling the order effect in the perception of paired stimuli

Hellström [7] accounts for the perceptual asymmetry in AX discrimination (i.e. same-different judgment) tasks by weighting the perception of the individual stimuli to be discriminated, as they are positioned in mental space in reference to a prototypical representation. The magnitude of the perceptual weights depends on a stimulus' position inside the stimulus pair :

$$d_{ab} = \left\{ \sum_{i=1}^{p} [s_1(\Psi_{ai} - \Psi_{ri}) - s_2(\Psi_{bi} - \Psi_{ri})]^m \right\}^{1/m} \quad s_1 < s_2 \tag{1}$$

[1]In the following article, the two terms will be used interchangeably

$$d_{ba} = \left\{ \sum_{i=1}^{p} \left[s_1(\Psi_{bi} - \Psi_{ri}) - s_2(\Psi_{ai} - \Psi_{ri}) \right]^m \right\}^{1/m} \quad s_1 < s_2 \tag{2}$$

where d_{ab} is the perceived distance between the two items of the stimulus pair a-b, and d_{ba} is the perceived distance between the same two items, presented in reverse order. Ψ are stimulus representations: representations with subscripts a and b refer to the memory trace of stimuli a and b. Ψ_r is the reference level, i.e. the reference point in the category against which other stimuli are compared. i to p indicates the number of dimensions. s_1 and s_2 are sensation weights, for which the first is assumed to be smaller than the second, thus attributing a smaller weight to the distance between the first stimulus and the prototype than to the distance between the second one and the prototype. m is the Minkowski constant [7].

Thus, when the first stimulus is closer to the reference level, the distance between the two stimuli appears smaller than when the second stimulus is closer to the reference level, albeit the acoustic difference remained constant. Furthermore should the asymmetries be larger for pairs in which both stimuli are fairly distant to the reference level, than for pairs in which one stimulus more or less coincides with the reference level; as the distance of both stimuli to the reference level is larger, the rescaling done by the sensation weights becomes more noticable.

The current study applies this reasoning to a reanalysis of a subset of the Blizzard 2008 data [8]. Mean opinion scores (MOS) are used to guide the direction of hypotheses for which order effects are expected in same-different judgments after classifying stimuli into two groups: high and low scorers. If we consider pairs of stimuli where one stimuli is picked from each group, we can make predictions:

The perceived distance between paired stimuli will be larger if the higher scoring stimulus is presented first compared to when it is presented second; within the high scoring stimuli group, which consists of natural speech examples and very good synthesized speech examples, the natural examples are still superior, and thus we can expect the order effect to be larger for comparisons involving the good synthetic examples than those of natural examples.

Figure 1 shows the expected relationships schematically: As the distance between natural stimuli and low ranking synthesized speech is larger than the distance between high and low ranked synthesized speech, comparisons of the former kind will receive more *different* judgments than comparisons of the latter kind. This should be generally true, regardless of the direction of comparison (see arrow lengths in Fig. 1).

More *different* judgments should also be found if the less natural stimulus is presented in second, rather than in first position, regardless of which kind of more natural stimulus is presented first (cf. lengths of paired arrows in opposing directions). However, this asymmetry should be more pronounced if both stimuli are synthesized than if one stimulus is natural (cf. relative differences in lengths of paired arrows).

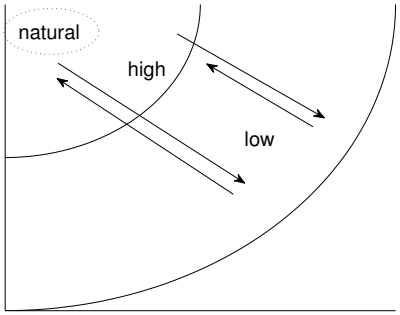

Figure 1: Hypotheses for the order effect in the perception of synthesized speech; the base of the arrow is the first stimulus in a pair, the tip is the second

3. Methods

We are re-analysing a dataset that was collected as part of a previous study [9] to address asymmetries between orders of presentation with a guided hypothesis. The dataset comprises of the evaluations from 30 native English speakers who produced judgements on 10 stimuli, 2 of which were natural speech (stimuli T1, B1) and 8 of which where synthetic (T2–T5, B2–B5) covering a wide range of naturalness.

For the set of stimuli, subjects provided both same-different judgements for each pair of stimuli in each presentation orders, and Mean Opinion Scores (MOS) for each of the stimuli. See [9] for the full experimental procedure.

4. Initial Re-analysis

Ordinal MDS plots (using the Identity Euclidian distance metric) were generated for the same-different responses, one for the lower, and one for the upper matrix triangle. The thus generated coordinates were overlayed in figure 2 with the corresponding T1 points aligned as anchors to get a first impression of whether order of presentation played any role in the data.

4.1. Results

Initial inspection of the MDS graphs, shows that there are some differences between the plots for each order.

When ranking the stimuli according to their distance to the natural stimulus T1, we already see that between the two configurations, the ranks of T3 and B5 are interchanged. This demonstrates that testing stimulus pairs in only one order of presentation can have strong effects

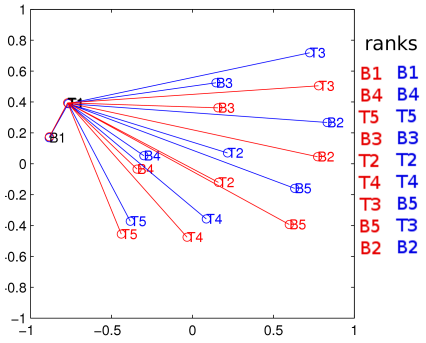

Figure 2: The two MDS configurations, once on lower, once on upper triangle of the data matrix, and their distance to T1

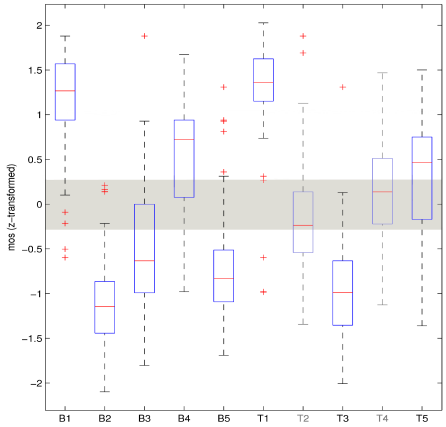

Figure 3: Boxplots of MOS scores collected for the 10 stimuli; grey boxes were not used in the asymmetry analysis

- even for such a low number of stimuli and despite the fact that we have a full matrix for every participant, and the same stimulus set has been tested on all listeners.

5. Further Analysis

MOSs are often recorded as conditional similarity data, which means that values cannot be compared directly between subjects [10]. In order to allow comparisons between listeners, MOS were transformed into z-scores for every listener and then averaged across participants. Thus relations between stimuli were maintained, even though the ranges of the rating scales participants used, differed.

In order to make clear predictions, the data were partitioned into two groups (see Figure 3):

- a high group: whose MOS were better than average. As a minimum, a z-score of 0.25 was selected.
- a low group: whose MOS were worse than average. As a maximum, a z-score of -0.25 was selected.

This constraint excludes two stimuli, T2 and T4, as their medians lie close to $z = 0$. This constraint also ensures that we have two distinct groups with no overlap. The high ranking stimuli T1, B1, B4, T5 were paired in turn with each of the low ranking stimuli B2, B3, B5, and T3 to test the hypothesis that pairs consisting of a member of the high ranking group presented first, and a member of the low ranking group second, will receive more *different* judgments than the same stimuli presented in the reverse order.

Additionally, since B4 and T5, the synthetic stimuli in the high ranking group, were rated as less natural than the natural example stimuli in this group, we hypothesise that the asymmetries found in pairs involving B4 and T5 should be larger than those comparisons which involve stimuli B1 and T1.

5.1. Results

To get a clearer picture of the effect order of presentation has, we test our hypothesis using a McNemar test (which tests the difference between paired judgments) on the individual same-different response data. This shows that the presentation order of the more natural-sounding stimulus inside a stimulus pair has a significant effect on the perceived similarity of a pair ($\chi^2(1)$=8.348, p<.01, the odds-ratio is 11.4). This indicates that the asymmetries we have found between the two MDS configurations above are systematic, rather than noise, determined by the position of the more typical items in stimulus pairs.

To investigate how the selected stimulus pairs contribute to this order effect, the trials were pooled across listeners for every contrast. To ensure that the asymmetry effect is constituted by several stimulus pairings, rather than one which is very pronounced, a Wilcoxon signed rank test is run[2]. It takes the proportions of *different* judgments in stimulus pairs as input and conducts pairwise comparisons between the two orders of presentation.

It revealed that there are significantly more *different* judgments in pairs in which the more natural sounding stimulus came in first (Mdn=27), as opposed to in second position (Mdn=26.5), T=24M p<.05, r=-0.516. This shows that the effect is not constituted by a single stimulus pair, but overall supports the asymmetry hypothesis.

Looking at the proportions of *different* judgments (see Figure 4), the highest bars are the ones in B1 and T1 comparisons, since the differences between these two natural stimuli and the synthesized sentences were largest. Blue bars show contrasts, in which the more natural stimulus was presented first, while the reverse is true for red bars. So according to the order effect we expect that blue bars are higher than red bars. Generalizing, this is true. However, while this effect is very pronounced for comparisons between synthesized sentences, this is not

[2] As the Shapiro-Wilk test for the data was significant for the variable of directional change towards the prototype, a non-parametric test had to be chosen

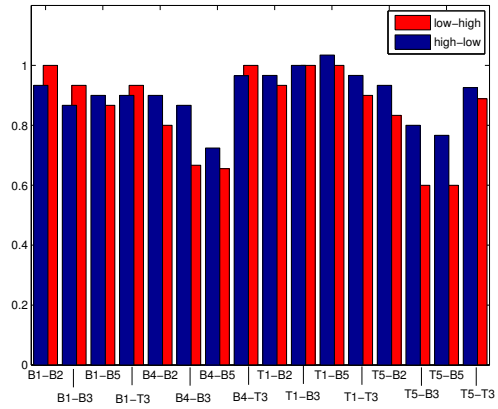

Figure 4: Number of *different* judgments for stimulus pairs across 30 listeners

indiscriminately true for natural stimuli: T1-B3 shows equal perception of naturalness, and in most B1 contrasts the effect is reversed! B1 comparisons clearly contradict the hypothesis. While we assumed that the order effect would be smaller for comparisons involving natural stimuli than only synthesized ones, we still expected an effect in the same direction. As for the order effects in more natural-sounding synthesized pairs B4 and T5, the order effect is present, and more pronounced than in the natural stimuli, which supports the model of the order effect.

6. Discussion

It is an interesting finding to see that the presentation order effect can be consistently found across listeners, rather than only in a repeated measurements in a within subjects design. This supports the assumption that native speakers of English have a fairly similar idea of what *natural* sounds like. This is the precondition for testing the naturalness of synthesized speech. The finding also stresses the need to gather discrimination data in *both* directions when conducting MDS analysis.

It is particularly interesting that while the order effect was demonstrated robustly for pairs including the high ranking synthesized examples, it was not found for pairs involving natural recordings, especially in the case of B1. Why our predictions do not hold true entirely for the recordings of a natural voice, we can only hypothesize at this point. It could be that we have insufficient data to see this effect clearly, or it could be that the natural speech has a different status to that of synthesized speech, which both results in the higher number of *different* responses and the loss of the order effect. It is also worth considering that these data were not originally collected with this kind of investigation in mind, and thus several parameters were left uncontrolled for: there was no control over the amount of time that elapsed between playing the first and the second stimulus of a pair, for which an order effect may be susceptible to [12]; neither were repeated plays of a stimulus recorded.

A question raised for future work is as to whether the order effect itself can be used to pin-point the location or direction of prototypical, natural speech in an MDS derived space where there were no natural stimuli included in the experiment.

7. Acknowledgements

The authors would like to thank The Blizzard Challenge organisers and participants for providing the data used for this work. The research leading to these results has received funding from the International Max Planck Research School Neuroscience of Communication and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287678.

8. References

[1] Janska, A.C., "Further Investigation of MDS as a Tool for Evaluation of Speech Quality of Synthesized Speech", MSc Dissertation, The University of Edinburgh, 2009. Online: www.era.lib.ed.ac.uk/handle/1842/3624

[2] Minda, J.P., Smith, J.D., "Prototype models of categorization: Basic formulation, predictions, and limitations", in Pothos, E.M., and Willis, A.J. [Eds], Formal Approaches in Categorization, 40–64, Cambridge University Press, 2011.

[3] Rosch, E., Mervis, C.B., "Family resemblances: Studies in the internal structure of categories", Cognitive Psychology, 7(4): 573–605, 1975.

[4] Rosch, E., "Cognitive Reference Points", Cognitive Psychology, 7(4): 532–47, 1975.

[5] Schellenberg, E. G., "Asymmetries in the discrimination of musical intervals: Going out-of-tune is more noticeable than going in-tune", Music Perception, 19(2): 223–248, 2001.

[6] Hellström, A., "Anatomy of stimulus comparison", in J.S. Monahan, S.M. Sheffert & J.T. Townsend [Eds], Fechner Day 2005, 113–118. Proceedings of the 21st Annual Meeting of the International Society for Psychophysics, Traverse City, Michigan, USA, October 19-22, 2005.

[7] Hellström, A, "Temporal asymmetry and "magnet effect" in similarity and discrimination of prototypical and nonprototypical stimuli: Consequences of differential sensation weighting", in S. Mori, T. Miyaoka, & W. Wong [Eds], Fechner Day 2007, 283–288. Proceedings of the 23rd Annual Meeting of the International Society for Psychophysics. Tokyo: International Society for Psychophysics, 2007.

[8] Karaiskos, V., King, S., Clark, R., Mayo, C., "The Blizzard Challenge 2008". URL: http://festvox.org/blizzard/bc2008/summary Blizzard2008.pdf, 2008

[9] Janska, A.C., Clark, R.A.J., "Further exploration of the possibilities and pitfalls of multidimensional scaling as a tool for the evaluation of the quality of synthesized speech". The 7th ISCA Tutorial and Research Workshop on Speech Synthesis: 142–147, 2010

[10] Coxon, A., Jackson, J.E., Davies, P.M., Smith H.V., Sachs, L., Schmee, J., "User's guide to multidimensional scaling", Heineman Education books, 1982.

[11] Janska, A.C, Clark, R.A.J., "Native and non-native speaker judgements on the quality of synthesized speech". Proc. Interspeech 2010: 1121–1124, 2010.

[12] Hellström, A., Rammsayer, T.H., "Effects of time-order, interstimulus interval, and feedback in duration discrimination of noise bursts in the 50- and 1000-ms ranges", Acta Psychologica, 116: 1–20, 2004.