Extrinsic normalization for vocal tracts depends on the signal, not on attention

Matthias J. Sjerps[1]*, James M. McQueen*[1,2] *& Holger Mitterer*[1]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2] Behavioural Science Institute and Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition, Radboud University Nijmegen, Nijmegen, The Netherlands
m.j.sjerps@gmail.com, j.mcqueen@pwo.ru.nl, holger.mitterer@mpi.nl

Abstract

When perceiving vowels, listeners adjust to speaker-specific vocal-tract characteristics (such as F_1) through "extrinsic vowel normalization". This effect is observed as a shift in the location of categorization boundaries of vowel continua. Similar effects have been found with non-speech. Non-speech materials, however, have consistently led to smaller effect-sizes, perhaps because of a lack of attention to non-speech. The present study investigated this possibility. Non-speech materials that had previously been shown to elicit reduced normalization effects were tested again, with the addition of an attention manipulation. The results show that increased attention does not lead to increased normalization effects, suggesting that vowel normalization is mainly determined by bottom-up signal characteristics.

Index terms: vowel normalization, speech perception, attention

1. Introduction

Listeners compensate for the voice characteristics of a speaker in a preceding sentence when listening to speech [1]. For example, when listeners categorize targets that lie on an F_1-vowel continuum, such as /pɪt/ - /pɛt/, they reveal a strong influence of the F_1 range in a preceding sentence [1]. That is, when the precursors are manipulated to have either a high or a low F_1 contour, normalization effects are observed as a shift in categorization of the following targets. Normalization has been argued to reflect listeners' ability to "tune in" to a particular speaker's vocal-tract properties [1]. This compensation mechanism has been argued to have a mainly auditory basis [2], [3]. For instance, when categorizing target vowels spoken by a male speaker, listeners compensate for the voice properties of a precursor even when this was spoken by a female [4]. Moreover, with spectrally rotated speech signals, normalization effects have been observed that are qualitatively similar to those obtained with speech signals (spectral rotation changes the frequencies of the formants but preserves the spectrotemporal complexity of the signal: however, these signals are generally interpreted as non-speech [5]). Furthermore, normalization effects with tone sequences as precursors have been reported [6] and for speech materials normalization effects have been observed at relatively early processing stages as reflected in electroencephalographic signals [7] (during the N1 time-window, which is associated with pre-categorical processing).

Recent investigations of normalization effects with non-speech materials, however, have questioned the purely auditory nature of normalization effects. [5] manipulated speech signals in a number of ways that made the new stimuli sound unlike speech, but that preserved the gross spectral similarity between the speech and the non-speech signals. First, speech stimuli were created with precursors that had a high and a low F_1, along with a target continuum, also defined by F_1 (i.e., /pɪt/ - /pɛt/). These were then both manipulated. One of these manipulations was spectral rotation. Spectrally rotating a speech sound causes the F_1 contour to be moved to a completely different frequency region (to keep the acoustic relation between the precursors and the targets similar, the targets were also spectrally rotated). Moreover, a number of other manipulations were applied to the precursor. The question was whether these manipulations would influence the normalization effects. [5] found that some of the non-speech precursors that were created did not induce normalization effects – precursors created, for instance, by removing low amplitude parts (such as silent closures in stops), by setting the pitch contour to a fixed value (at ~224 Hz), by temporally reversing the syllables and setting them to an equal amplitude, and by spectrally rotating them around 1250 Hz.

Crucially, the findings reported in [5] thus show that normalization processes do not always occur. Variation in the strength of normalization, especially with non-speech materials, has been reported on a number of occasions (see [8] and references therein). A potential explanation is that the discrepancy between normalization with speech signals versus no normalization with non-speech signals reflects an attentional effect. In particular, it could reflect an influence of how relevant listeners judge the precursor signal to be in relation to the perception of the following target. Based on perceived relevance, listeners may pay more or less attention to the precursors. Because speech signals are naturally more informative than non-speech signals, it is likely that listeners will pay more attention to speech than to non-speech stimuli. Lack of normalization with non-speech signals may thus reflect listeners' lack of attention to those signals. In fact, if this alternative explanation, focusing on attention as a latent variable, were correct, this criticism would not only apply to the current study but to many more designs where non-speech materials led to effects that differed from those with speech signals [3]. Alternatively, however, the strength or occurrence of normalization could depend on specific auditory properties of the precursor (apart from the Long Term Average Spectrum, or LTAS, relation between precursor and target because this aspect was matched in [5]). The current experiments were set up to test the hypothesis that the reduction of normalization effects with non-speech materials came about because the non-speech materials that had previously been used did not capture attention as much as speech materials did.

To investigate the contribution of attention to extrinsic vowel normalization, two experiments were run. These experiments made use of the materials reported in [5], but introduced an additional task that focused participants' attention on the precursors. This allowed for a direct comparison between the effect sizes obtained in the current experiment with an attentional manipulation with those in the earlier data without an attentional manipulation. In some of the earlier experiments, [5] reported no normalization, but two types of manipulated materials *did* produce normalization effects. Those were used in the current experiments. The first type of material was spectrally rotated speech with no additional manipulations. The second set of materials to lead to normalization effects were speech targets, which were preceded by a precursor that was manipulated by removing its low amplitude parts, setting its pitch contour to a fixed value, temporally reversing the syllables and setting them to an equal amplitude (i.e., in the same ways as the most extreme manipulation described above, *except* that there was no spectral rotation). For the targets the pitch was also set at a fixed value in order to create similarity between the precursor and target. The choice for the materials that gave rise to normalization effects was made because signals that have been shown to induce small effects will probably be more susceptible to an attentional manipulation than those previously showing null effects. This is, this strategy prevents a potential floor effect.

The procedure was very similar to that in the experiments reported in [5]. Participants were asked to categorize spectrally-rotated speech targets (Experiment 1) or speech targets that had a flat pitch (Experiment 2). These targets were preceded by precursors. Within an experiment, a precursor could have either a high F_1 or a low F_1 (or spectrally-rotated analogs of these formant values in Experiment 1). Additionally, the precursors in Experiment 2 were manipulated in a number of ways (no low amplitude parts, flat pitch contour, reversed syllables of equal amplitude; see below for further details). Critically, and in contrast to [5], an additional task encouraged participants to attend to the precursors. In this additional task, participants were asked to refrain from responding to the target whenever the precursor had a dip in amplitude (these catch-trial precursors were presented occasionally throughout the experiment). In summary, the main goal of this study was to test the influence of an attentional task on the size of vowel normalization effects.

2. Experiments

2.1. Participants

Twenty native speakers of Dutch were recruited (12 in Experiment 1 and 8 in Experiment 2). Experiment 1 required 4 additional participants because there was considerable individual variation in effect size for the first 8, although the average effect was of similar magnitude and in the same direction as the average after 12.

2.2. Stimuli

2.2.1. Experiment 1

Base target sounds consisted of a six step [pɪt] to [pɛt] continuum (an F_1 distinction). The steps were created by lowering the F_1 from a recorded instance of /ɛ/. This instance was spoken by a native female speaker of Dutch. The average F_1 value of this instance of [ɛ] was 575 Hz, the average F_2 value was 1844 Hz (F_2 was not manipulated). To create a test continuum, the vocalic portion of the recording of the word /pɛt/ was excised. Using a Linear Predictive Coding (LPC) procedure, the source model (a model of the sound emitted from the vocal folds) was separated from the filter model (a model of the filter characteristics of the vocal tract) using 20 predictors. Using fewer predictors left remnants of the formants in the source model, which would have made it more difficult to shift the perceived identity of the targets towards /ɪ/. The formant filter model was based on 4 formants. The continuum was created by a linear decrease of F_1 over 200 Hz in 6 steps of 40 Hz. The formant and filter model were recombined to create the target vowel continuum. All materials were band-pass filtered between 200 and 2500 Hz. All targets were adjusted so that their overall amplitude and their amplitude envelope matched those of the original vowel instance of /pɛt/. The targets were then spectrally rotated around 1250 Hz. The precursors were based on a Dutch sentence (*"Op dat boek staat niet de naam*", lit. on that book is not the name). This sentence was manipulated, with the same procedure as for the targets, to have either a low F_1 (-200 Hz) or a high F_1 (+200 Hz). These versions of the precursor were then spectrally rotated in the same way as the targets.

2.2.2. Experiment 2

Target materials were the same as in Experiment 1, with the exception that they were *not* spectrally rotated and that they had a flat pitch level to increase similarity between the precursors and targets. The pitch was flattened using the overlap-add method for resynthesis in Praat [9]. The base speech precursors (i.e., the versions prior to spectral rotation) from Experiment 1 were used for Experiment 2. Several additional manipulations were applied to both the low- F_1 and the high- F_1 precursors. The signals were modified to have a flat pitch at the average value of the speech materials (223.8 Hz) using the same method as was used for the targets. Each of these signals was divided in high and low amplitude parts, leading to 6 high amplitude parts, roughly corresponding to the words. All the high amplitude parts were temporally reversed (e.g., the first digital sample of a part became the last sample of the new "reversed part" and vice versa) and equalized in amplitude relative to each other. All low amplitude parts (silences between words or those due to stop-closures) were excised and discarded.

2.3. Procedure

The training and testing procedures were identical in both experiments and were also identical to those used in [5] except for the addition of the catch trials for the attentional task (for a description see below).

2.3.1. Training

As the participants for Experiment 1 were presented with novel non-speech stimuli, they first had to undergo a three-phase training protocol to familiarize them with these materials. The same procedure was applied for Experiment 2 to keep the amount of exposure across the experiments similar. In each training phase, participants had to reach a performance criterion to go on to the next training phase or, after the third training phase, to the testing part of the experiment. During all three training phases, but not at test, visual feedback ("correct" (correct) or "fout" (incorrect)) appeared on a computer screen after each trial. The first training phase consisted of a discrimination task using only the endpoint targets (a same-different task; criterion: for three consecutive blocks, seven out of eight correct). The second phase consisted of a categorization task with the endpoint targets (with the options "A" and "B"; criterion: for three consecutive blocks, nine out of ten correct). Participants were told that they had to find out which target belonged to which button ("A" or "B"). The third training phase consisted of a categorization task that was similar to the second phase, with the addition that the targets were preceded by a neutral version of the precursor (i.e., a version with no F_1 manipulation).

During the third training phase there were also catch trials that indicated to participants that they should refrain from responding. These catch trials were not included in [5]. Catch trials could be recognized by a two-word long dip in amplitude of 20 dB. Catch trials (pseudo) randomly varied in where the amplitude dip would occur (2nd and 3rd word; 3rd and 4th word; 4th and 5th word; 5th, 6th and 7th word. The last pair consists of three words because it includes the article "de"). In order not to change the criterion relative to [5], erroneous button presses on the catch trials did not influence whether participants could pass to the test phase. One catch trial was presented every 10 trials.

2.3.2. *Testing*

In both experiments, the six target steps were each played after both the high and low precursors (in random order) for 15 repetitions, resulting in 180 test trials (with two self-paced pauses). Trials were presented without feedback. Participants categorized the targets with the same two buttons as those used during the second and third training phases ("A" and "B"). In addition to the test trials, the testing phase also contained catch trials that were constructed in the same way as those in the last training phase. On every block of twelve trials (6 steps x 2 precursors) two additional catch trials (one with a high F_1 and one with a low F_1) were presented. After such precursors, the middlemost step of the continuum was presented (halfway between steps 3 and 4). As in the training phase, participants had to refrain from responding on catch trials. All stimuli were presented with a 500 ms silent interval between the precursor and the following target.

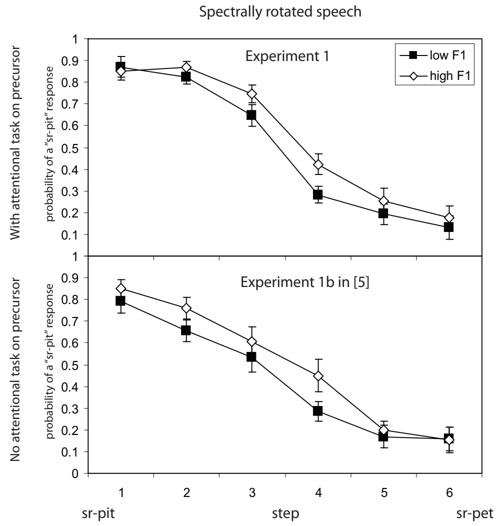

Figure 1: *Probability of spectrally rotated /pɪt/ ("sr-pit") responses to stimuli on a sr-/pɪt/ to sr-/pɛt/ continuum. Targets were preceded by precursors that were manipulated to have a high or a low F_1 and that were also spectrally rotated. Top panel: data for results reported here, in a task where participants were encouraged to pay attention to the precursors. Bottom panel: data for results reported in [5] with no attentional task. Error bars reflect standard errors of the mean.*

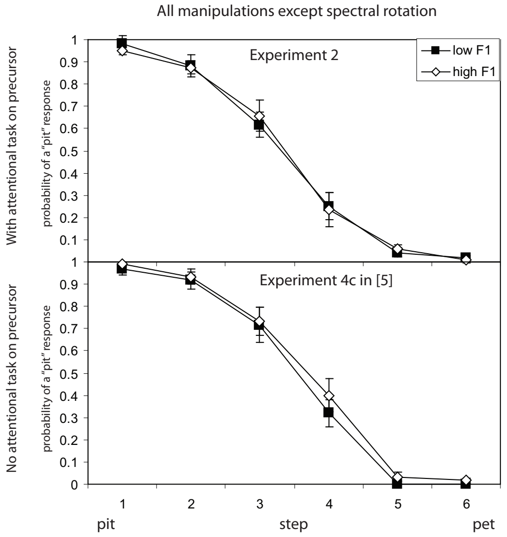

Figure 2: *Probability of /pɪt/ ("pit") responses to stimuli on a /pɪt/ to /pɛt/ continuum. Targets were preceded by precursors that were manipulated to have a high or a low F_1. They were further manipulated to have a flat pitch, no low-amplitude parts and temporally reversed syllables that had equal amplitudes. Top panel: data for results reported here, in a task where participants were encouraged to pay attention to the precursors. Bottom panel: data for results reported in [5] with no attentional task. Error bars reflect standard errors of the mean.*

2.4. Results

The data were analyzed using linear mixed-effects regression models in R (version 2.6.2, R development core team, 2008, with the lmer function from the lme4 package of [10]). Different models were tested in a backward elimination procedure, starting from a complete model. All factors were numerical and centered around 0. These included the factors Step (levels: -2.5 to 2.5 in steps of 1), Precursor (levels: low F_1 = -1 vs. high F_1 = 1), Block (15 stimulus repetitions: levels -7 to 7 in steps of 1), Attention (data reported in [5]: -1; data reported in current paper: 1) and their possible interactions. For fixed factors, non-significant predictors were taken out of each analysis in a stepwise fashion, starting from the highest order interaction, until no predictors could be removed without significant loss of fit. If an interaction was only just significant, the optimal model without this interaction was also found and then the two models were compared by means of a likelihood ratio test. A full random-effects structure was implemented (involving by-participant slopes for Block, Precursor, Step, and their interactions).

2.4.1. *Experiment 1*

On seventy-two percent of the catch trials participants correctly refrained from responding. Only the data for the non-catch trials were further analyzed. The top panel of Figure 1 displays the results obtained here, the lower panel of Figure 1 displays the results from the corresponding experiment in [5] (reported there as Experiment 1b). They revealed effects in the same, compensatory direction. In the comparison of the effects obtained here and in [5], modeling settled on main effects for the factors Step (b_{Step} = -0.903, p < 0.001, reflecting the steepness of the categorization curve) and Precursor ($b_{Precursor}$ = 0.200, p = 0.001, reflecting the contrastive normalization effect). There was no interaction between Precursor and

Attention, indicating that the addition of the attentional task had no effect on the strength of normalization. A Chi-square goodness of fit comparison between the optimal model, and the same model, but with the addition of the interaction between Attention and Precursor revealed that inclusion of the interaction was not warranted ($df = 2$, $\chi^2 = 1.49$, $p = 0.475$).

2.4.2. Experiment 2

On sixty-two percent of the catch trials participants correctly refrained from responding. Analyses were carried out on the non-catch trial data. The top panel of Figure 2 displays the results obtained here, with no visible compensation effect. The bottom panel of Figure 2 displays the results from Experiment 4c in [5], for which a small effect in the compensatory direction was reported. In an overall analysis comparing the two datasets, the optimal model settled on a single main effect of Step ($b_{Step} = -2.200$, $p < 0.001$, reflecting the steepness of the categorization curves), with no further interactions. A comparison between this model and one for which the interaction between Attention and Precursor was added (along with their main effects) showed that the model without the interaction was optimal ($df = 3$, $\chi^2 = 3.562$, $p = 0.313$).

3. Discussion

This study investigated the role of attention on extrinsic vowel normalization. Two experiments reported in [5] were replicated here with an additional attentional task. Listeners categorized speech targets on a [pɪt] to [pɛt] continuum (Experiment 2) or spectrally rotated versions of these (Experiment 1). These targets were preceded by precursors, in two conditions, manipulated to have an F_1 contour that was increased or decreased by 200 Hz. In Experiment 1, these precursors were also spectrally rotated. In Experiment 2, the precursors were not spectrally rotated but manipulated in a number of other ways (no low amplitude parts, flat pitch contour, reversed syllables of equal amplitude). Furthermore, in contrast to [5], listeners in both experiments were encouraged to pay attention to the precursors through the inclusion of an additional task: they had to refrain from responding when a dip in amplitude occurred in the precursor.

The attentional manipulation did not increase the amount of normalization. A normalization effect was found in Experiment 1 and this effect was of a similar size to the effect in [5]. Although no significant normalization effect was found in Experiment 2, the difference between this experiment and that in [5] was not significant. It appears that paying attention to a precursor sound does not change the precursor's influence on categorization of subsequent sounds (and Experiment 2 suggests that, if anything, attention might make the effect smaller).

Normalization effects have been argued to reflect a biological solution to optimize information processing in changing environments [12]. Despite advances in our understanding of these compensatory mechanisms, and especially the similarity between normalization with speech and non-speech signals [2], [3], [4], [5], [6], [12], the fact that effects obtained with non-speech signals are generally smaller remains understudied. In the introduction we formulated two potential explanations for the difference in effect-sizes between normalization with speech and non-speech signals. The first explanation, tested here, was that reduced effects with non-speech were the result of a reduction in attention to the precursor materials. The current study allows this explanation to be rejected. The alternative is that low-level acoustic properties of the signals led to the difference between speech and non-speech. But what is it, then, about those low-level acoustic properties that makes normalization effects stronger with speech signals? We propose that, through experience, listeners have acquired knowledge about the low-level characteristics of speech signals. This includes the knowledge that it is beneficial to compensate for the acoustic properties of the source. When listening to a given speaker such a process is highly beneficial because overall voice properties of speakers remain relatively stable. For other signals there may be less acoustic stability. Perceptual learning at very low levels of processing has been reported, for instance in the domain of speech pitch perception [11]. We suggest that the perceptual system has gained experience with the spectrotemporal characteristics of speech and has learnt to process subsequent signals with similar spectrotemporal complexity (possibly captured by prosody) in a non-independent fashion. This can partly be done by taking the spectral characteristics of preceding contexts into account [12]. It is noteworthy that [5] reported that rated "speechiness" of context signals did not predict the size of normalization. This means that although signals are required to be similar to speech for normalization processes to occur, the similarity to speech is determined based on bottom-up signal characteristics, and not directly related to subjective impressions.

To conclude, the current study indicates that differences in effect-sizes in extrinsic normalization between speech and non-speech signals are not because speech captures more attention than non-speech. The size of normalization effects and indeed their very occurrence appear to be determined mainly by bottom-up signal properties.

4. References

[1] Ladefoged, P., and Broadbent, D. E. "Information conveyed by vowels". Journal of the Acoustical Society of America, 29, 98-104, 1957.

[2] Watkins, A. J., and Makin, S. J. "Perceptual compensation for speaker differences and for spectral-envelope distortion". Journal of the Acoustical Society of America, 96, 1263-1282, 1994.

[3] Watkins, A. J., and Makin, S. J. "Effects of spectral contrast on perceptual compensation for spectral-envelope distortion". Journal of the Acoustical Society of America, 99, 3749-3757. 1996.

[4] Watkins, A. J. "Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion". Journal of the Acoustical Society of America, 90, 2942-2955, 1991.

[5] Sjerps, M. J., Mitterer, H., and McQueen, J. M. "Constraints on the processes responsible for the extrinsic normalization of vowels". Attention , Perception and Psychophysics, 73, 1195–1215, 2011.

[6] Holt, L. L. "Temporally nonadjacent nonlinguistic sounds affect speech categorization". Psychological Science, 16, 305-312, 2005.

[7] Sjerps, M. J., Mitterer, H., & McQueen, J. M. "Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics". Neuropsychologia, 49, 3831– 3846, 2011.

[8] Sjerps, M. J., Mitterer, H., & McQueen, J. M. "Hemispheric differences in the effects of context on vowel perception". Brain and Language, 120, 401–405, 2012.

[9] Boersma, P., and Weenink, D. Praat: Doing phonetics by computer [Computer software]. http://www.praat.org, 2005.

[10] Bates, D. M., and Sarkar, D. lme4: Linear mixed-effects models using S4 classes, R package version 0.99875-6, 2007.

[11] Krishnan, A., Xu, Y., Gandour, J., and Cariani, P. "Encoding of pitch in the human brainstem is sensitive to language experience". Cognitive Brain Research, 25, 161–168, 2005.

[12] Kluender, K. R., and Kiefte, M. J. "Speech perception within a biologically realistic information-theoretic framework". In M. A. Gernsbacher and M. Traxler [Eds], Handbook of Psycholinguistics (2nd ed., 153-199). Elsevier, 2006.