Atten Percept Psychophys
DOI 10.3758/s13414-011-0096-8

Constraints on the processes responsible for the extrinsic normalization of vowels

Matthias J. Sjerps · Holger Mitterer · James M. McQueen

Abstract Listeners tune in to talkers' vowels through extrinsic normalization. We asked here whether this process could be based on compensation for the long-term average spectrum (LTAS) of preceding sounds and whether the mechanisms responsible for normalization are indifferent to the nature of those sounds. If so, normalization should apply to nonspeech stimuli. Previous findings were replicated with first-formant ($F1$) manipulations of speech. Targets on a [pɪt]–[pɛt] (low–high $F1$) continuum were labeled as [pɪt] more after high-$F1$ than after low-$F1$ precursors. Spectrally rotated nonspeech versions of these materials produced similar normalization. None occurred, however, with nonspeech stimuli that were less speechlike, even though precursor–target LTAS relations were equivalent to those used earlier. Additional experiments investigated the roles of pitch movement, amplitude variation, formant location, and the stimuli's perceived similarity to speech. It appears that normalization is not restricted to speech but that the nature of the preceding sounds does matter. Extrinsic normalization of vowels is due, at least in part, to an auditory process that may require familiarity with the spectrotemporal characteristics of speech.

Keywords Speech perception · Psycholinguistics · Extrinsic vowel normalization · Speech vs. nonspeech · Compensation

M. J. Sjerps (✉) · H. Mitterer · J. M. McQueen
Max Planck Institute for Psycholinguistics,
P.O.box 310, 6500 AH Nijmegen, The Netherlands
e-mail: matthias.sjerps@mpi.nl

J. M. McQueen
Behavioural Science Institute and Donders Institute for Brain, Cognition and Behaviour, Centre for Cognition,
Radboud University Nijmegen,
P.O.box 9104, 6500 HE Nijmegen, The Netherlands

Our interpretation of auditory events is dependent on the context in which they occur. Context-dependent interpretation helps listeners resolve speech sound ambiguities such as those that arise from speaker differences. For example, the interpretation of vowels depends on the first- and second-formant characteristics of the speaker who utters those vowels (Ladefoged & Broadbent, 1957). One account of this process proposes that it is the result of a general auditory mechanism that normalizes perception of any acoustic input by constructing a long-term average of the distribution of frequencies of a sound source (Kiefte & Kluender, 2008; Watkins & Makin, 1996). On the basis of this long-term average, the perceptual impact of acoustic energy in certain frequency regions of a subsequent target sound becomes attenuated. This mechanism would thus influence the interpretation of speech sounds on the basis of spectral information in the preceding sentence (or extrinsic context). According to this account, the same adjustments should apply to nonspeech target sounds following nonspeech precursors. The present article investigates normalization of speech and versions of those stimuli that have undergone extensive manipulations to make them unlike speech. It was thus tested whether extrinsic normalization could be based solely on this general auditory long-term average spectrum (LTAS) compensation mechanism. It was also tested whether compensation mechanisms can operate independently of the acoustic and perceptual characteristics of the precursor signal.

The key finding behind the LTAS mechanism is that suppressing or exciting particular frequency regions of a precursor sentence can induce a shift in the categorization of subsequent vowels. Watkins (1991) found effects of normalization when listeners categorized targets on an /ɪ/-to-/ɛ/ continuum that were presented after intelligible precursor sentences that had been filtered. The filter suppressed either those frequencies that were generally

Published online: 15 February 2011

more pronounced in an instance of /ɛ/, as compared with /ɪ/ (i.e., an /ɪ/-minus-/ɛ/ filter), or the reverse (i.e., an /ɛ/-minus-/ɪ/ filter). Participants gave fewer /ɪ/ responses to targets after a precursor that was filtered with the /ɪ/-minus-/ɛ/ filter than after a precursor that was filtered with the /ɛ/-minus-/ɪ/ filter. Categorization appeared to be shifted because listeners were more sensitive to spectral properties that were suppressed in the precursor sentence, increasing the probability that the vowel was perceived as the one that had more energy in that frequency region. Watkins and Makin (1994, 1996) thus suggested that normalization occurs through a process that could be described as applying an inverse form of the precursor's average filter characteristics to the target sound.

Similarly, Kiefte and Kluender (2008) tested participants' /i/ versus /u/ categorizations with speech sounds from a 7 × 7 grid varying the second formant (*F*2) and spectral tilt (both of these aspects can influence perceived /i/ vs. /u/ identity; Kiefte & Kluender, 2005). When a precursor sentence had been processed by the same acoustic filter as that used to adjust the spectral tilt of the target stimuli, listeners relied only on the target *F*2 value. However, when the precursors had been filtered to match the *F*2 peak of the following vowel, listeners' /i/ versus /u/ responses were based on the target's spectral tilt alone. This effect shows that listeners suppress the information value of acoustic aspects that are invariant between a precursor and a subsequent target. The result is that listeners become less sensitive to static information but gain sensitivity for acoustic change. Normalization for a signal's long-term frequency characteristics does just that.

Watkins and Makin (1994) found that normalization can also be observed with filtered precursor sentences that are played backward and that the amount of normalization is not reduced in this case. Furthermore, Stilp, Alexander, Kiefte, and Kluender (2010) reported normalization effects on the perception of musical instruments. They created a target range from *saxophone* to *French horn*. Participants had to categorize these targets, which were presented after precursor sounds that were filtered to emphasize the spectral characteristics of either the French horn or the saxophone. They found—in analogy to previous findings with speech materials (e.g., Watkins & Makin, 1994, 1996)—that the categorization of the musical nonspeech targets shifted depending on the spectral characteristics of the precursor signal. These shifts were also of a contrastive nature and were observed with speech and instrumental precursors.

The account emerging from this body of prior research is that extrinsic normalization in speech perception is based on a general-purpose auditory mechanism that compensates for the LTAS characteristics of preceding speech and is indifferent to the nature of the sounds from which the LTAS is derived. However, Watkins and Makin (1994) also found that the amount of normalization was strongly reduced when a noise precursor was used, even though it had the same LTAS as the speech precursor. Moreover, Watkins and Makin (1991) found that the normalization effect can even be completely abolished when such noise precursors and the targets are presented to opposite ears. This finding led them to argue that normalization effects take place at at least two different stages: first, an initial peripheral stage that operates only over short intervals and with ipsilateral presentation; second, a central normalization stage that operates over longer precursor–target intervals and applies to stimuli that are presented ipsilaterally and contralaterally. They argued that the small amount of normalization that was found with the noise precursor with bilateral presentation was therefore completely due to peripheral auditory compensation effects. Watkins and Makin (1996) also suggested that a prerequisite for normalization at central processing stages might be that a precursor signal needs to contain spectrotemporal variation.

Our focus in the present study was on normalization at central processing levels. We asked whether the central compensation mechanism is based solely on the LTAS and, hence, whether it is completely indifferent to the exact nature of the precursor signal. The suggestion of Watkins and Makin (1996) that spectrotemporal variation is a prerequisite for normalization effects to occur at central processing levels already casts doubt on such a pure LTAS mechanism, but the process may have other prerequisites. There could be signals, acoustically intermediate between speech signals and signal-correlated noise, that also fail to induce central normalization effects. If so, this would raise the question of why certain signals induce compensatory effects, whereas others do not. In such a situation, a learning account would potentially offer a solution. Because adult listeners have had an abundance of exposure to speech from different speakers, they will gain experience with the fact that certain voice properties are stable within a speaker. They could therefore learn that it is beneficial to perceive vowels relative to those voice properties. This can be achieved if listeners learn to normalize for the LTAS properties of preceding sound sequences. Additionally, however, it would then be valuable for speech perception if central normalization was in some way restricted in order to avoid normalization for the wrong precursor signals. This means that it should apply only to signals with particular characteristics. We tested this learning hypothesis by manipulating the properties of precursor signals.

We began by testing whether compensation for the LTAS of a precursor signal is independent of the acoustic and perceptual characteristics of a precursor signal by comparing normalization at central processing levels in speech and nonspeech signals. Nonspeech signals were created by spectrally rotating speech sounds. Spectral rotation is a transformation that rotates the spectral makeup of a complex signal around a central frequency, such that the information in the high-frequency ranges trades places with the information in the low-frequency ranges. This transformation changes the frequencies of the formants but preserves the spectrotemporal complexity of the signal. Spectral rotation also keeps part of the pitch information of speech sounds intact (through the repetition rate in the waveform; Moore, 2003), although cues to pitch trough relations of the harmonic are lost. Spectral rotation also inverts the spectral tilt. That is, whereas voiced speech sounds invariably have less energy at higher frequencies, spectrally rotated versions have more energy at higher frequencies. If normalization is specific to speech sounds, it should not occur with spectrally rotated versions of speech. The general auditory account, however, predicts that spectral rotation should not prevent normalization, because spectral rotation of both the target and precursors keeps intact the overlap between the spectra.

In subsequent experiments we varied the auditory properties of speech and nonspeech precursor signals. According to the general auditory account, both speech and nonspeech signals should induce normalization effects through compensation for LTAS. According to an extension of the general auditory account that includes learning, the signal may need to have speechlike spectrotemporal characteristics, and thus compensation for LTAS would not be completely independent of more fine-grained temporal and spectral properties of the precursor signal.

Experiment 1

Experiment 1 tested the influence of speech precursors on speech sounds (Experiment 1a) and the influence of spectrally rotated precursors on spectrally rotated targets (Experiment 1b). In Experiment 1a, participants categorized targets on a [pɪt]-to-[pɛt] continuum (an *F*1 distinction), presented after speech precursors with an increased or a decreased average *F*1 level. In Experiment 1b, participants heard nonspeech stimuli that were created by spectrally rotating the materials of Experiment 1a (both precursor and target stimuli).

Since the participants in Experiment 1b were presented with novel nonspeech stimuli, they first had to undergo a three-part training protocol to familiarize them with the materials. Participants in Experiment 1a underwent the same training with the speech materials to ensure that the amount and type of stimulus exposure were similar over the two experiments.

There are two potential problems involved in using speech sounds and their spectrally rotated counterparts. The first stems from the fact that auditory frequency resolution decreases going from low frequencies to higher frequencies (Moore, 2003). Differences that would be audible in speech stimuli could become inaudible after spectral rotation. All materials were therefore low-pass filtered at 2.5 kHz and spectrally rotated around a frequency of 1.25 kHz.

While this minimizes the differences in spectral resolution between the original and the spectrally rotated materials, participants may still find it easier to discriminate between extremely familiar speech sounds than between their unfamiliar spectrally rotated counterparts. The discriminability of the material sets was therefore equated via a pretest, in which participants heard the speech and nonspeech stimuli in a staircase discrimination task (Appendix 1). Using the results from this pretest, speech and nonspeech stimuli were selected that differed by similar perceptual distances.

The second potential problem with speech and spectrally rotated speech materials is that they may induce the type of peripheral normalization effects argued to be distinct from more central processes (Watkins, 1991). These effects, which operate over short interstimulus intervals (ISIs), result in a shift in categorization functions in the same direction as normalization effects that take place at the more central levels of processing that are under investigation here. In order to prevent such peripheral effects with our materials, a precursor–target interval of 500 ms was used. This should thus ensure that any effect found is a true instance of longer term normalization.

Experiment 1a, which involved speech materials, was expected to result in clear normalization effects. It should thus provide a replication of numerous previous studies (Ladefoged & Broadbent, 1957; Watkins, 1991; Watkins & Makin, 1994, 1996). This effect would be characterized by a shift in categorization functions for targets presented after a speaker with a generally high *F*1 versus a speaker with a generally low *F*1. The results for Experiment 1b, however, depended on the nature of the compensation process. No categorization shift should occur if extrinsic normalization is restricted to speech signals. But if normalization is a process that is not restricted to intelligible speech, we would expect to observe a categorization shift similar to that predicted for Experiment 1a.

Method

Participants

Participants from the Max Planck Institute for Psycholinguistics participant pool were recruited and tested until 16 (8 for each part) had successfully completed the training and testing parts of the experiment (see Appendix 2 for details on all experiments and subexperiments). All were native speakers of Dutch, reported no hearing or language disorders, and had not participated in a similar experiment before. They received payment for their participation.

Materials

Experiment 1a: Targets All recordings were made by a female native speaker of Dutch. The materials were down-sampled offline to 11050 Hz. Acoustic processing of the stimuli was carried out using Praat software (Boersma & Weenink, 2005). The test syllables were the Dutch words /pɪt/ (the stone of a fruit) and /pɛt/ (cap). To create a test continuum, the vocalic portion of a 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 toward /ɪ/. The formant filter model was based on four formants. The continuum was created by a linear decrease of $F1$ over 200 Hz in steps of 1 Hz in the formant model. The formant and filter model were recombined to create the target vowel continuum. The average $F1$ value of the endpoint [ɛ] was 575 Hz; the average $F2$ value was 1844 Hz ($F2$ was not manipulated). The $F1$ values are close to the average $F1$ values found in female speakers of northern standard Dutch, which is closest to the dialect of the speaker (/ɛ/ = 535 Hz; /ɪ/ = 399 Hz; values from Adank, van Hout, & Smits, 2004), while the $F2$ value is relatively low (/ɛ/ = 1990 Hz; /ɪ/ = 2276 Hz). The manipulated vocalic portions were spliced back into the unmanipulated consonant context from [pɛt]. 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/.

On the basis of the pretest (Appendix 1), six target steps from the initial continuum were selected, ranging from [pɛt] to [pɪt]. These steps spanned an $F1$ range of 60 Hz in steps of 12 Hz (with $F1$ values ranging from 70 to 130 Hz lower than the recorded /pɛt/). The pretest showed that this frequency range resulted in clearly discriminable endpoints.

Experiment 1a: Precursors The precursor was the Dutch sentence "op dat boek staat niet de naam" [ɔp dɑt buk stat nit də nam] (on that book, it doesn't say the name), which contains, among others, the vowels [u], [a], [i], and [ə], thereby providing listeners with all of the point-vowels and schwa. The precursor did not contain the target vowels /ɪ/ and /ɛ/, to prevent direct precursor–target comparison. The average $F1$ value over the vocalic portions of the selected precursor sentence was 502 Hz, ranging from roughly 300 Hz (in [i]) to 800 Hz (in [a]). The $F1$ values of the vocalic portions of the precursors were, in two versions, either increased or decreased by 200 Hz (after Watkins & Makin, 1994), using the same method as that which was used for the target vowels. The formant filter model was based on four formants, except for the vowel portion of the word /nit/, which was based on three formants. Along with the surrounding vowels, the first nasal of the word /nam/ was also included in the manipulation, since that increased the naturalness of the token. The other two nasals were unmanipulated in both conditions. The manipulated vocalic portions were spliced back into the unmanipulated consonantal parts of the original precursor sentence.

Experiment 1b: Targets The targets were created in the same way as those in Experiment 1a, with the addition of the critical manipulation that the signals were spectrally rotated around 1250 Hz. The pretest (Appendix 1) determined that the 60-Hz $F1$ difference used for the speech targets was too small to be detected for the spectrally rotated versions of the targets. The $F1$ difference between the endpoints for Experiment 1b was therefore set at the full range of 200 Hz, leading to approximately equally discriminable test stimuli across the experiments. A six-step continuum between the two endpoints was selected (steps of 40 Hz).

Experiment 1b: Precursors The precursors that were used in Experiment 1a were spectrally rotated around 1250 Hz (the same manipulation as that applied to the targets in Experiment 1b).

LTAS measures Figures 1 and 2 display the LTAS (bin width = 10 Hz) of each precursor and each endpoint target, along with each difference LTAS, for both subexperiments. The x-axis is logarithmic. Figure 1 shows that the low-$F1$ precursor has more energy than does the high-$F1$ precursor at low frequencies. Although, for example, the low-$F1$ precursor LTAS is not perfectly matched to the LTAS of the targets, the difference lines show that the relative differences between the low-$F1$ and the high-$F1$ stimuli are indeed matched to the differences between the target endpoints. This means that those frequencies that listeners use to distinguish between /ɛ/ and /ɪ/ are roughly the same frequencies as those that constitute the acoustic difference between the precursors in the high- and low-$F1$ conditions. Both precursors have more energy at higher frequencies than do the targets (the target spectral tilt has a steeper slope). This should thus, in principle, cause an ambiguous target to be perceptually slightly shifted toward /pɪt/ for both precursor conditions. However, the difference line between the precursors shows that the induced shifts toward /pɪt/ will be stronger for the high-$F1$ condition, since it has more energy at the higher frequencies than does the low-$F1$ precursor. The two precursor conditions were therefore predicted to induce different target categorization functions. The focus was therefore on the relative differences among the precursors matched to the relative difference among the targets.[1]

The comparison between Figs. 1 and 2 shows that the relation between the difference spectra for the precursors and their respective endpoint targets was similar across the two subexperiments. The high-$F1$ speech precursor had more energy at frequencies above the average $F1$ value than did the low-$F1$ speech precursor, and the /pɛt/ endpoint had more energy at frequencies above the average $F1$ frequency than did the /pɪt/ endpoint. Spectral rotation reversed these differences but preserved the similarities in the relation between the precursors and targets. The spectrally rotated precursor based on the high-$F1$ speech precursor thus had more energy at frequencies below the average spectrally rotated $F1$ than did the low-$F1$ spectrally rotated precursor. This was also the case for the spectrally rotated /pɛt/ endpoint, relative to the spectrally rotated /pɪt/ endpoint.

Design and procedure

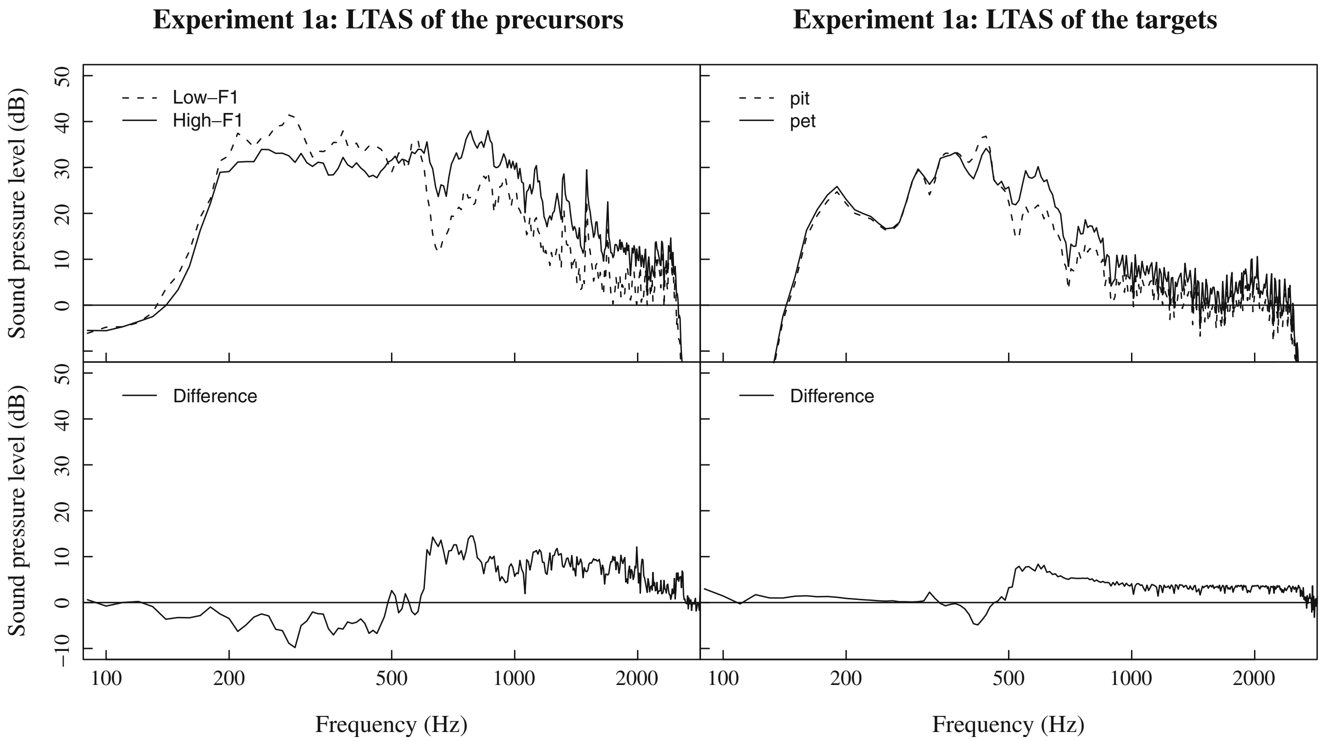
The experiment was run using Presentation software (Version 11.3, Neurobehavioural Systems Inc.). All auditory stimuli were presented binaurally, through Sennheiser HD 280-13 headphones.

Training In the three-part training procedure, used in both subexperiments, participants learned to categorize the two unambiguous endpoint stimuli. The first part consisted of a discrimination experiment. On every trial, participants heard a combination of the two endpoints (words in Experiment 1a; spectrally rotated versions of these words in Experiment 1b). The task was to indicate, by pressing a button, whether the two stimuli heard on a trial were the same or different. Visual feedback ("correct” [correct] or "fout” [incorrect]) appeared on a computer screen after each trial. If participants had seven out of eight responses correct on three consecutive blocks, they entered the second part of the training. For this second part, participants listened to the same two endpoint stimuli, but in isolation. They were told that it was their task to find out which stimulus belonged to which button label ("A" or "B"), using the feedback they would receive, and that they would thus initially have to guess. Visual feedback was the same as in the first part. If participants reached a 90% correct criterion over three blocks (of 10 stimuli each), they entered the third part. This again involved categorizing these same two sounds as "A" or "B" with feedback, but the sounds were now presented after an unmanipulated (neutral) version of the precursor that was to be used in the testing phase. The same criterion applied as in the second part. Within all three parts, participants were allowed a self-paced pause after every 100 trials.

Testing In each subexperiment, the six target steps were each played after the two precursors for 15 repetitions, resulting in 180 test trials (with two pauses). This phase took about 12 min. Trials were presented without feedback. Participants categorized the targets by means of the same two buttons as those used during the second and third training phases ("A" and "B").

Data analysis In this and all the following experiments, responses faster than 100 ms after target onset were excluded. Luce (1986) showed that simple responses to auditory stimuli start from 100 to 150 ms after stimulus onset. Any faster responses could thus not have been due to the perception of the target stimuli. After exclusion of missed responses and responses that were too fast, 99.7% of the trials were kept, on average, over Experiments 1, 2, 3, and 4 (the lowest proportion of preserved responses was 99.5%; no fast responses needed to be excluded in Experiment 5). The data were analyzed using linear mixed-effects models in R (Version 2.6.2 [R development core team, 2008], with the lmer function from the lme4 package of Bates & Sarkar, 2007). Different models were tested in a backward elimination procedure, starting from a complete model. All factors were entered as numerical variables, centered around 0. These included the factors step (level on the continuum: -2.5 to 2.5 in steps of 1), precursor (levels; low $F1$ [-1] vs. high $F1$ [1]), and block (15 stimulus repetitions: levels -7 to 7 in steps of 1) and all their possible interactions. Nonsignificant predictors were taken out of the analysis in a stepwise fashion, starting from the highest order interaction, until no predictors could be

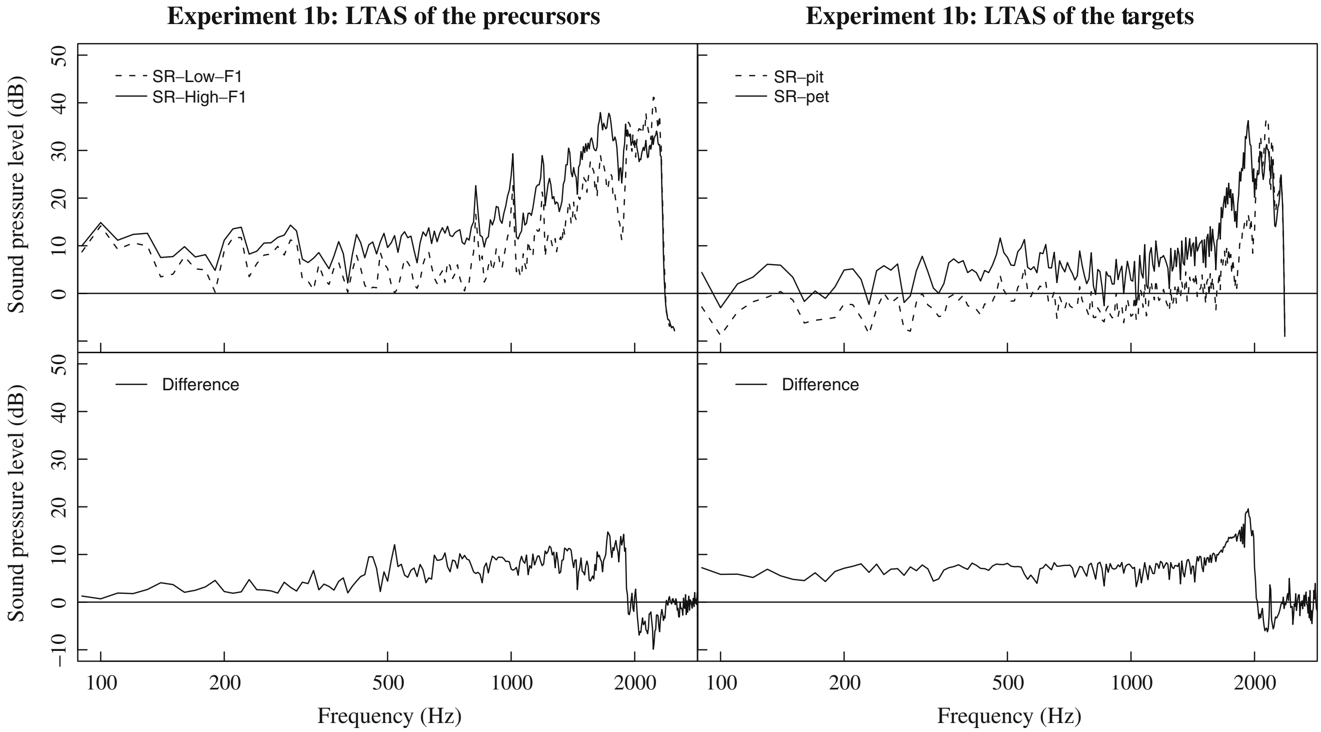
[1] This discussion raises the question of whether it would be possible to model the predicted direction and amount of normalization that should be observed with any combination of precursor and target. Although this is beyond the scope of the present article, it would, in principle, be possible. A model would have to include a number of parameters. First, one would have to specify which frequency components are perceptually important for the distinction between two different vowels and to what extent. Second, a model would probably have to restrict the influence of any precursor frequency component to only target frequency components that have overlapping tuning curves. The size of these tuning curves will be dependent on the level of processing that is modeled (peripheral or central), since these might not be of an equal size. Third, since the contextual influence of a certain acoustic event is probably restricted to some amount of time, it is likely that the most recent part of a precursor signal has more influence on the direction and amount of compensation than do parts of a precursor signal that are most distant (and also, here the amount differs depending on the level of processing that is modeled). Thus, this would require additional weighting functions for temporal distance.



Experiment 1a: LTAS of the precursors

Experiment 1a: LTAS of the targets

Fig. 1 Long-term average spectrum (LTAS) plots for the speech materials in Experiment 1a. Upper left panel: LTAS for the high-*F*1 precursor (solid line) and for the low-*F*1 precursor (dotted line). Bottom left panel: The difference spectrum for the precursors. Upper right panel: LTAS for each endpoint target, /pɛt/ (solid line) and /pɪt/ (dotted line). Bottom right panel: The difference spectrum for the targets



Experiment 1b: LTAS of the precursors

Experiment 1b: LTAS of the targets

Fig. 2 Long-term average spectrum (LTAS) plots for the spectrally rotated materials in Experiment 1b. Upper left panel: LTAS of the spectrally rotated (SR) versions of the precursors with a high *F*1 (solid line) and with a low *F*1 (dotted line). Bottom left panel: The difference spectrum for the precursors. Upper right panel: LTAS for each SR endpoint target /pɛt/ (solid line) or /pɪt/ (dotted line). Bottom right panel: The difference spectrum for the targets

removed without significant loss of fit. If an interaction was only just significant, the optimal model without this interaction was also found. The best model was then established by means of a likelihood ratio test.

Results

Experiment 1a The upper panel of Fig. 3 shows the average categorization results. The optimal model for the data had significant main effects of the factor step, b = -0.777, p < .001, which indicates the decrease of /pɪt/ responses toward the /pɛt/ end of the continuum; block, b = -0.030, p = .046, which indicates that the number of /pɪt/ responses decreased as the experiment progressed; and precursor, b = 0.556, p < .001. The latter effect indicates that the probability of a /pɪt/ (low $F1$) response is much higher after high- than after low-$F1$ precursors. Interaction effects were found between the factors step and block, b = 0.033, p = .001, indicating that participants' categorizations became, overall, less categorical as the experiment progressed; and between the factors block and precursor, b = -0.040, p = .010, reflecting the fact that the effect of precursor became smaller toward the end of the experiment, although it was never absent.

Experiment 1b The bottom panel of Fig. 3 displays the results. The optimal model had main effects of the factor step, b = -0.726, p < .001, which indicates that the probability of a “nonspeech /pɪt/” response decreases when moving along the continuum from “nonspeech /pɪt/” to “nonspeech /pɛt/"; and an effect of the factor precursor, b = 0.196, p = .002. The effect of precursor indicates that the probability of a "nonspeech /pɪt/” (NS-/pɪt/) response is higher for targets after a spectrally rotated high-$F1$ precursor than for those after a spectrally rotated low-$F1$ precursor. This is an effect in the same direction as that for the speech stimuli.

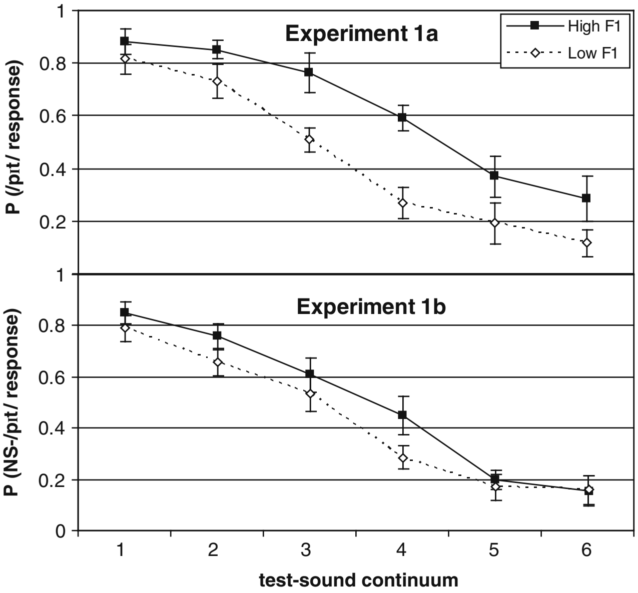


Fig. 3 Experiments 1a (upper panel) and 1b (lower panel): Probability of a /pɪt/ or “nonspeech /pɪt/” (NS-/pɪt/) response to a continuum of the targets /pɪt/ (step 1) and /pɛt/ (step 6) in Experiment 1a or the spectrally rotated analogs of these in Experiment 1b. Targets were presented after precursor sentences with an $F1$ value that was increased (high $F1$) or decreased (low $F1$) by 200 Hz (or the spectrally rotated analogs for Experiment 1b). Error bars reflect standard errors

Discussion

Experiment 1a showed that the spectral properties of a precursor sentence can influence the categorization of a vowel continuum. This replicated earlier findings (e.g., Ladefoged & Broadbent, 1957; Watkins & Makin, 1994). Experiment 1b showed that this finding is not restricted to the processing of speech sounds. Spectrally rotated versions of the stimuli used in Experiment 1a resulted in a normalization effect that was similar to that found in Experiment 1a. The size of this effect was reduced in Experiment 1b, but this probably reflects the fact that the difference in frequency between the endpoint stimuli was much larger for the spectrally rotated stimuli (60 Hz for speech, 200 Hz for the spectrally rotated speech). The step size for the spectrally rotated targets in Experiment 1b was much larger (steps of 40 Hz) than the step size for the speech targets in Experiment 1a (steps of 12 Hz). Watkins and Makin (1996) have demonstrated that the ratio of the spectral contrast over the target continuum to that of the precursor continuum has a strong influence on the size of the categorization shifts. With respect to our stimuli, the difference spectrum for the targets of Experiment 1a was smaller than the difference spectrum for the targets in Experiment 1b, whereas the difference spectra for the precursors in Experiments 1a and 1b were of the same size. Given the findings by Watkins and Makin (1996), this could have resulted in the smaller normalization effect in Experiment 1b. As is evident from the results of the pretest, a larger target step size was necessary for participants to categorize the nonspeech target range reliably. The focus here is thus on the qualitative finding that normalization occurred with unintelligible stimuli, and over a relatively long precursor–target interval.

Two additional observations from the results of Experiment 1a deserve to be discussed. The first is that in Experiment 1a, the effect of precursor decreased as the experiment progressed. This was probably due to the fact that the effect of a precursor potentially extends beyond the trial on which it is presented. Since the different precursor conditions were presented intermixed, listeners could have been increasingly influenced by precursors from both conditions. It is unclear why this effect did not occur in Experiment 1b, however. The second observation is that the proportion of /pɪt/ responses decreased as the experiment progressed. It is unclear what the exact nature of this effect is, but it is possible that listeners started the test phase with a slight bias toward the /pɪt/ end of the continuum and compensated for this bias as the experiment progressed (Repp & Liberman, 1987).

In sum, however, the results of Experiment 1 support the LTAS compensation account of extrinsic normalization. Moreover, the auditory compensation mechanism seems to apply to both speech and nonspeech materials, suggesting that the central processing mechanism is general and independent of the acoustic and perceptual aspects of the precursor (apart from its LTAS). This conclusion seems to conflict with the finding by Watkins (1991) that, with noise precursors, no effects of precursors are found. One might therefore argue that the materials in Experiment 1b were too much like speech and, hence, that the results could be explained by a speech-specific mechanism. Although spectral rotation destroyed the phonetic content of the original sentence, the precursors used in Experiment 1b still had many speechlike prosodic characteristics. In Experiment 2, the materials were manipulated in ways that still preserved spectrotemporal variation in the precursors. More important, it also preserved the LTAS relation between the precursors and the targets. However, the manipulations rendered the materials acoustically more unlike speech (as compared with the materials in Experiment 1b). These manipulations consisted of: removing pitch variation, removing the (very) low-amplitude parts (e.g., low-amplitude parts of stop closures), temporally reversing the individual syllables, equalizing the average amplitudes of the syllables, and, again, spectral rotation of all the materials. If normalization at a central processing level is the result of a general auditory process that compensates for LTAS, completely independently of the acoustic nature of the precursor, these extremely nonspeechlike materials should still induce a normalization effect. If Experiment 2 did not result in a normalization effect, however, this would suggest that there are, in fact, restrictions to the type of precursor sounds that can induce normalization.

Experiment 2

Method

Participants

Participants from the Max Planck Institute for Psycholinguistics participant pool were recruited and tested until 8 had successfully completed the training and testing parts of the experiment (see Appendix 2). Participants received a monetary reward for their participation. None had participated in Experiment 1.

Materials

Targets The targets were the same as those in Experiment 1b. But to maintain similarity to the new precursors, they were now manipulated such that they had a flat pitch level (at the average level of the original target). These pitch adjustments were made using the overlap-add method for resynthesis in Praat.

Precursors The precursors that had been used in Experiment 1a (low $F1$, unmanipulated, high $F1$) were first modified to have a flat pitch at the average value of the speech materials (223.8 Hz), using the same method as that used for the targets. Each of these signals was then divided into high- and low-amplitude parts (see the upper panel of Fig. 4; non-annotated sections are considered low amplitude). All the high-amplitude parts were temporally reversed (e.g., the first digital sample of a word became the last sample of the new reversed word and vice versa) and were equalized in amplitude, relative to each other. Reversing only the individual syllables, rather than reversing the complete sentence, has the advantage that the pattern of LTAS change as the sentence develops is very similar across the speech and nonspeech stimuli (and much more similar than would be the case if the complete sentence were reversed). All low-amplitude parts were excised and discarded. The resulting signals were then spectrally rotated. The bottom panel of Fig. 4 displays the resulting precursor and target in one condition. The manipulations that were applied in this and the following experiments are summarized in Table 1.

LTAS measures Figure 5 displays the LTAS plots of the precursors and the endpoint targets, along with their difference LTAS. If compared with the LTAS plots in Experiment 1b (see Fig. 2), the present LTAS plots may seem quite different at first glance, with reasonably smooth spectra in Fig. 2 and peaky spectra in Fig. 5. This is a consequence of the pitch-flattening procedure; the peaks represent the harmonics of the constant $f0$. If one focuses on the difference between the spectra of high-$F1$ and low-$F1$ versions, however, the relations are similar across Experiments 1 and 2. The nonspeech high-$F1$ precursor had more energy at frequencies below the average $F1$ value than did the low-$F1$ precursor. Similarly, the nonspeech-/pɛt/ endpoint had more energy at frequencies below the average $F1$ frequency than did the nonspeech-/pɪt/ endpoint.

Fig. 4 Upper panel: Experiment 1a. Annotated spectrogram of the original precursor sentence in the low-$F1$ condition followed by /pɛt/. Bottom panel: Experiment 2. Annotated spectrogram of the manipulated precursor in the low-$F1$ condition followed by nonspeech /pɛt/. The symbol “~” indicates that the materials were manipulated

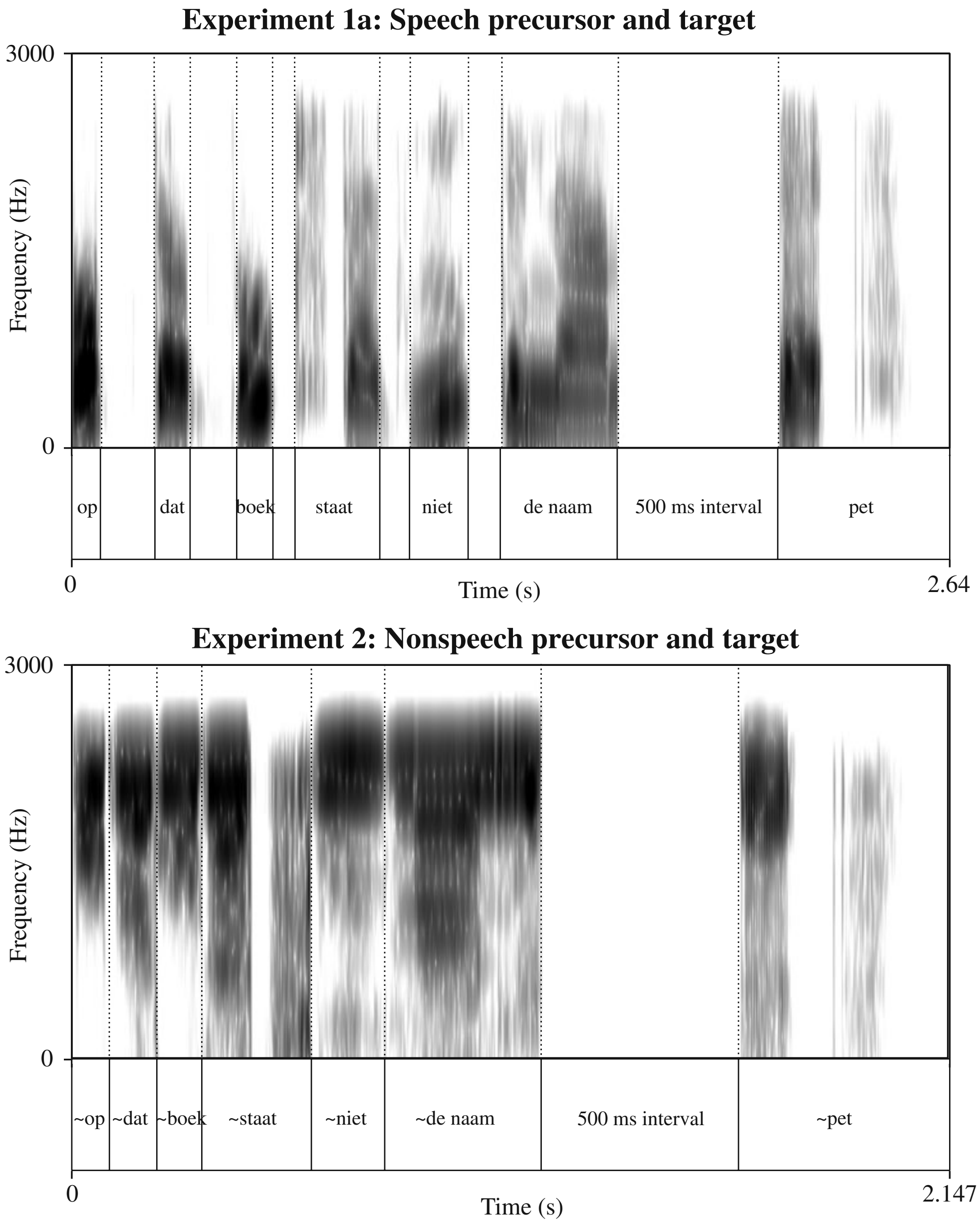


Table 1 Summary of precursor manipulations and results

Experiment	Precursor manipulation					Normalization effect found
	Pitch flattening	Reversed syllables	Spectral rotation	Equal amplitude	Breaks removed	
1a	No	No	No	No	No	Yes
1b	No	No	Yes	No	No	Yes
2	Yes	Yes	Yes	Yes	Yes	No
4a	No	Yes	Yes	Yes	Yes	No
4b	Yes	Yes	Yes	Yes	No	No
4c & 4d	Yes	Yes	No	Yes	Yes	Yes

Precursor sentences were or were not manipulated with respect to several speech characteristics: pitch movement, reversal of syllables, spectral rotation, syllables of equal amplitude, presence versus absence of low amplitude parts (breaks)

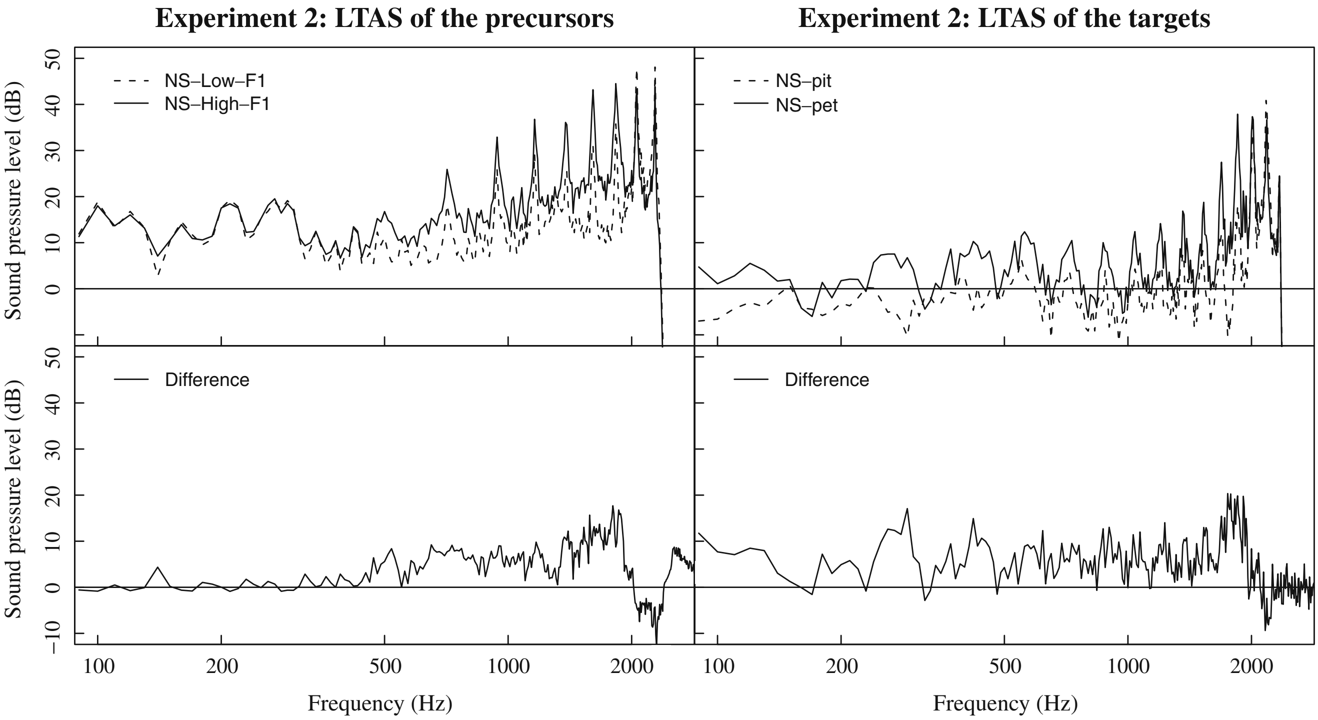


Fig. 5 Long-term average spectrum (LTAS) plots for the nonspeech materials in Experiment 2. Upper left panels: LTAS for the nonspeech sounds that originated from precursors with a high $F1$ (solid line) and a low $F1$ (dotted line). Left bottom panels: The difference spectrum for the nonspeech precursors. Upper right panels: LTAS for each nonspeech endpoint target that originated from the speech sound /pɛt/ (solid line) or /pɪt/ (dotted line). Right bottom panels: The difference spectrum for the nonspeech targets

Results

Figure 6 displays the results. Modeling settled on main effects for the factors step, b = -0.945, p < .001, indicating robust categorization, and block, b = -0.057, p < .001, which reflects a drift toward fewer overall nonspeech-/pɪt/ (NS-/pɪt/) responses during the experiment. An interaction was found between the factors step and block, b = 0.040, p < .001, which shows that responses became less categorical as the experiment progressed. When the precursor factor was included, it did not have a significant effect, b = -0.116, p = .084. There was only a trend in the direction opposite from that observed in Experiments 1a and 1b, and this trend was limited to one step.

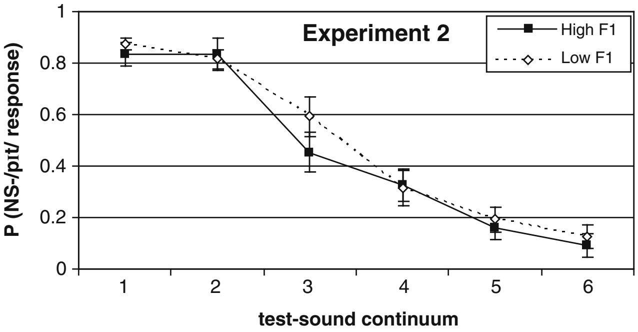


Fig. 6 Experiment 2: Probability of “nonspeech /pɪt/” (NS-/pɪt/) responses to a continuum of nonspeech targets that were manipulated versions of a range from /pɪt/ (step 1) to /pɛt/ (step 6), presented after manipulated versions of the precursor sentences used in Experiment 1a. Error bars reflect standard errors

Additionally, a comparison was made between the effects in Experiment 1b and the effects in Experiment 2, to compare the effects obtained with these two nonspeech experiments. The experiment factor was included in the analysis (levels; Experiment 1b [-1] vs. Experiment 2 [1]). The optimal model showed main effects for the factors step, b = -0.836, p < .001, and block, b = -0.025, p = .019. Two-way interactions were found between the factors experiment and step, b = -0.104, p = .001), such that responses in Experiment 2 were more categorical than those in Experiment 1b; between experiment and block, b = -0.032, p = .003, such that there was a decrease in the number of nonspeech /pɪt/ responses only in Experiment 2; and between experiment and precursor, b = -0.157, p < .001. The latter effect reflects the critical comparison between the effects of precursor over Experiments 1b and 2. The effect shows that there was significantly more normalization with the materials in Experiment 1b. A three-way interaction was found between the factors experiment, step, and block, b = 0.029, p < .001, indicating that only in

Experiment 2 did responses become less categorical as the experiment progressed.

Discussion

In contrast to the normalization effects obtained in Experiments 1a and 1b, Experiment 2 failed to show a normalization effect for nonspeech sounds. A comparison between the experiments showed that the size of the effect in the context condition was statistically different. An absence of compensation for LTAS was found in Experiment 2, despite a precursor–target LTAS relation that was very similar to that in Experiment 1b. As a comparison between the bottom panels of Figs. 2 and 5 shows, the precursor difference LTAS overlaps with the target difference LTAS for both experiments in a similar way. If extrinsic normalization had been the result of compensation for precursor LTAS, independently of the exact nature of the precursor, it should also have been found in Experiment 2.

It thus appears that a general LTAS compensation account that assumes indifference to the nature of the precursor signal is inadequate for centrally located compensation processes. However, one could argue that the presence of normalization in Experiment 1b, but not in Experiment 2, was due to acoustic differences between the stimuli. It is possible that the acoustic manipulations in Experiment 2 led to some subtle change in the relation of the LTAS of the precursors and targets, not apparent from the difference spectra, that prevented the normalization effect. To investigate this possibility, a control experiment was set up. This experiment was designed to establish what the effect would be of the LTAS of the precursors used in Experiments 1b and 2 if it was physically applied to their respective target stimuli. The target sounds from each experiment (Experiment 3a for the Experiment 1b stimuli and Experiment 3b for those in Experiment 2) were filtered such that those frequencies that were most pronounced in the precursor sentences would be most suppressed in the target signals. This method physically influences the target sounds in the same way as the hypothesized mechanism for LTAS compensation (cf. Watkins, 1991), in analogy to the approach taken in Watkins and Makin (1994). Since the precursors in Experiment 1b elicited a significant shift in target categorization, it was predicted that Experiment 3a would result in a strong shift in categorization. If the lack of a categorization shift in Experiment 2 was due to insufficient overlap between the difference LTAS of the precursors and targets, Experiment 3b should not show a shift in categorization. If, however, there is sufficient overlap between the difference LTAS of the Experiment 2 precursors and targets, we should find a shift in categorization in Experiment 3b that is similar to that predicted in Experiment 3a.

Experiment 3

Method

Participants

Eight further participants from the Max Planck Institute for Psycholinguistics participant pool were recruited and tested. They received a monetary reward for their participation.

Materials

The training materials in Experiments 3a and 3b were the endpoint stimuli that were used as training and test materials in Experiments 1b and 2, respectively. The test materials in Experiments 3a and 3b were filtered versions of the test materials used in Experiments 1b and 2. To create these stimuli, the LTAS of each precursor from the earlier experiment was applied as an inverse filter to all steps of the target continua. This means that the amplitude for every frequency was attenuated by the relative average amplitude of the precursor signal at that frequency. Those frequencies that were most pronounced in the precursor signals were thus also relatively most suppressed in the new target sounds. This operation mimics the situation that perception of the frequency distribution of a target sound is perceived relative to the frequency distribution of the precursor to its full extent. Since the manipulation resulted in signals with very low overall amplitudes, the amplitudes of all the targets were increased by 20 dB (equally across the whole spectrum), such that participants would be able to listen to these new targets at a comfortable hearing level. Figure 7 displays the LTAS of the resulting nonspeech /pɛt/ target sounds filtered by both the high-$F1$ and the low-$F1$ precursors from Experiment 2 (only the LTAS for one step of the continuum is displayed, because the difference spectra are the same for all steps). A comparison of the resulting difference spectrum with the difference spectrum of the precursors from Experiment 2 (Fig. 5, left bottom panel) shows that they are indeed the same. This means that we successfully applied the inverse LTAS filters of the precursors to the target continua. The same manipulation was applied to the targets in Experiment 1b (filtered by their appropriate precursor).

The same participants took part in both subexperiments. Half the participants were trained and tested on the Experiment 3a materials first and then on the Experiment 3b materials; for the other participants, this order was reversed. Training in the first subexperiment (3a or 3b) consisted of the first two phases from the earlier experiments (discrimination and then categorization), using the identical endpoint stimuli as those used in either Experi-

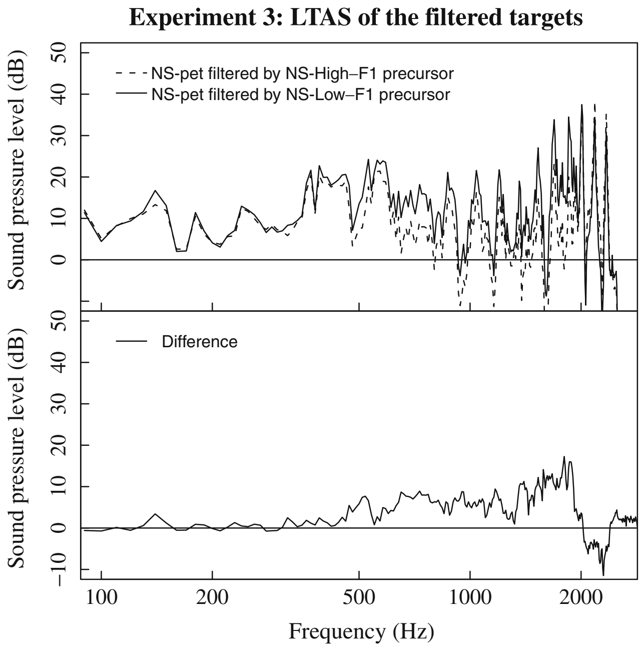


Fig. 7 Upper panel: Long-term average spectrum (LTAS) plots for the endpoint "nonspeech /pɛt/" targets taken from Experiment 2, attenuated by a filter that was constructed from the low-$F1$ precursor (solid line) and the high-$F1$ precursor (dotted line) taken from Experiment 2. Bottom panel: The difference spectrum for the resulting two Experiment 3 sounds. Whereas, for the other figures, the difference spectrum were calculated by subtracting the low-$F1$ variant from the high-$F1$ variant, here the subtraction was made in the opposite direction for ease of comparison

ment 1b (for Experiment 3a) or Experiment 2 (for Experiment 3b). The first testing phase consisted only of the new precursor LTAS filtered target continua, coming from the appropriate earlier experiment. In the second subexperiment, only the second training phase (categorization) was presented (the stimuli differed only in pitch characteristics and, thus, were already familiar to the participants). This second training phase was followed by the second test phase (again, with the appropriate precursor LTAS attenuated target continua).

Results

Experiment 3a The top panel of Fig. 8 displays the results. The order of the two parts was added as the factor order (levels; first: [-1] vs. second [1]) in the data analysis. The optimal model showed main effects for the factors precursor filter (i.e., which precursor from the earlier experiment was used to filter the target), b = 2.307, p < .001, and step, b = -0.879, p < .001. Two-way interactions were found between the factors step and order, b = 0.117, p = .009, indicating fewer categorical responses if this part was presented as the second one; and between order and block, b = -0.056, p = .003, indicating fewer nonspeech /pɪt/ responses toward the end of this part of the experiment.

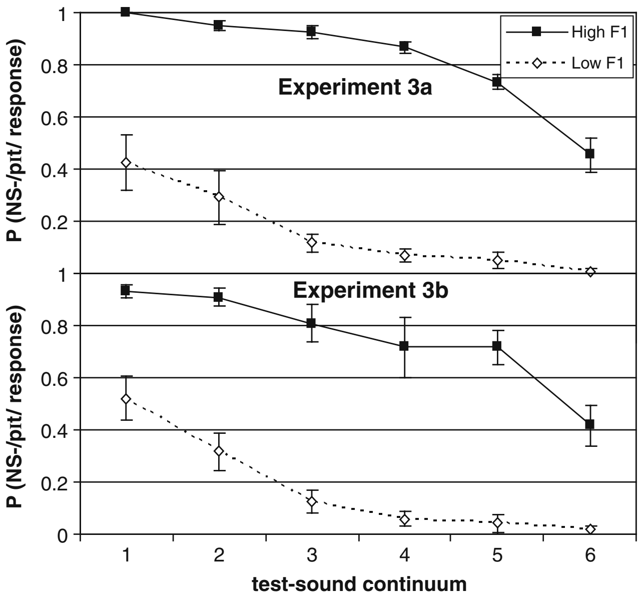


Fig. 8 Control experiments: Experiments 3a (upper panel) and 3b (bottom panel). The graphs display the probability of "nonspeech /pɪt/" (NS-/pɪt/) responses to the continuum of nonspeech targets from Experiment 1b (upper panel) or Experiment 2 (bottom panel), ranging from "nonspeech /pɪt/" (step 1) to "nonspeech /pɛt/" (step 6). The targets were attenuated by the filter properties of the LTAS of the nonspeech high-$F1$ or low-$F1$ precursors that were used in Experiment 1b (upper panel) or Experiment 2 (bottom panel). Error bars reflect standard errors

Experiment 3b The bottom panel of Fig. 8 displays the results. The optimal model showed main effects for the factors precursor filter, b = 1.856, p < .001, and step, b = -0.762, p < .001. Two-way interactions were found between the factors precursor filter and step, b = 0.158, p = .005, indicating a stronger effect of precursor filter toward the "nonspeech /pɛt/" end of the continuum; and between precursor filter and order, b = -0.281, p < .001, indicating a smaller effect of precursor filter if this part was presented as the second one.

Experiment 3a versus 3b An additional comparison was made between the effects obtained in the two subexperiments. This analysis included the factor experiment, which modeled the difference between the two subexperiments (levels; 3a: [-1] vs. 3b [1]). The optimal model showed an effect for precursor filter, b = 2.096, p < .001, step, b = -0.833, p < 0.001; and experiment, b = -0.215, p = .004, indicating fewer overall nonspeech /pɪt/ responses in Experiment 3b. Two-way interactions were found between experiment and precursor filter, b = -0.215, p = .005, indicating a smaller effect for precursor filter in Experiment 3b; between precursor filter and order, b = -0.221, p = .004, indicating a smaller effect of precursor filter if this subexperiment was presented second; and between step and order ,b = 0.112, p = .008, indicating fewer categorical responses if this part was presented as the second one. Three-way interactions were found between the factors precursor filter, step, and experiment, b = 0.117, p = .006, reflecting the fact that the interaction between precursor filter and step was not present in Experiment 3a; between precursor-filter, block, and order, b = -0.035, p = .005, indicating a progressive decrease in the effect of precursor filter, but only if this subexperiment was presented second; and between precursor filter, experiment, and order, b = -0.168, p = .002, reflecting that the difference in size of the effect of precursor filter across Experiments 3a and 3b was present only when subexperiments that were presented second were compared.

Discussion

Both Experiments 3a and 3b resulted in large categorization shifts. The spectral relation between the precursors and targets was thus similar across Experiments 1 and 2; that is, these control experiments demonstrate that the LTAS relation between the precursors and targets in both earlier experiments was such that a compensatory influence from the precursors should result in a contrast effect in both cases. The lack of a normalization effect in Experiment 2 was thus apparently not due to an insufficient match of spectral properties between the precursors and targets in those materials. It appears, instead, that the properties of the precursors in Experiment 2 were not appropriate for central normalization to take place.

It should be noted that the compensation effect in Experiment 3a (derived from the spectrally rotated speech in Experiment 1b) was bigger than the compensation effect in Experiment 3b (derived from the more extreme nonspeech materials in Experiment 2), and this was the case only when subexperiments that were presented second were compared. While we have no straightforward interpretation of this pattern, it is unlikely that this small difference in the size of the effect in this control experiment could explain the large difference between the results of Experiments 1b (compensation effect) and Experiment 2 (no compensation effect, with a trend in the opposite direction).

Since Experiment 2 showed that not all precursor signals influence perception of subsequent targets in the same way, an LTAS compensation process that is indifferent to the exact nature of the precursors cannot fully account for whether normalization effects occur. This raises the question of which type of precursors give rise to compensation for LTAS. There are a number of differences between the precursors used in Experiment 1b (which elicited compensation effects) and Experiment 2 (which failed to elicit compensation effects). Any one of these acoustic aspects might account for the absence of effects in Experiment 2. Experiment 4 addressed this issue by testing whether the presence of pitch variation in the precursor signals is a crucial factor (Experiment 4a), whether the presence of high- and low-amplitude parts in the signal can induce normalization processes (Experiment 4b), and whether a speechlike spectral tilt plays a critical role (Experiments 4c and 4d). In natural speech, these aspects are, to some degree, almost always present in the signal. In the materials for Experiment 2, however, they were removed. If any of them are necessary for extrinsic normalization to occur, one or more subexperiments should reveal a normalization effect.

Experiment 4

Method

Participants

For each of the four subexperiments, 8 different participants were recruited from the Max Planck Institute for Psycholinguistics participant pool (see Appendix 2). Participants received a monetary reward for their participation. None had taken part in the earlier experiments.

Materials

For Experiments 4a and 4b, the targets were identical to those used in Experiment 2 (spectrally rotated speech with a flat pitch contour). The precursors of Experiment 4a were created by imposing a sinusoid pitch contour (formula = 223.8 + 79.1 * sin [t * (3.14/(1.78/2))]) onto the precursors in Experiment 2. The contour had two periods and a pitch range of a size similar to the range of the pitch contour in the speech precursors. The overall amplitude was set at the same level as for the other experiments. For Experiment 4b, the materials were created by reintroducing zero-amplitude parts into the precursors in Experiment 2 at the same locations as the low-amplitude parts in the speech version (Fig. 4a). However, these parts were all of equal duration (0.107 ms), which was the average duration of the low-amplitude parts in the original speech materials. The materials of Experiment 4c (both precursors and targets) consisted of those from Experiment 2, but then spectrally rotated back, such that low frequencies became high frequencies and vice versa. The result of this operation is that the formants reappeared in their original locations and the signal regained a speechlike spectral tilt.

It would, however, be interesting to compare the results of Experiment 1a (speech that had all the speechlike characteristics) with the results of Experiment

4c (speech that was manipulated such that some of the speechlike characteristics were removed). This would be difficult, however, since the target range was different between Experiments 1a and 4c. To allow for this comparison, an additional experiment was run. Experiment 4d used the same precursor as Experiment 4c, but with the target sounds from Experiment 1a that had an $F1$ range of 60 Hz.

Results

Experiment 4a: Pitch movement The top left panel of Fig. 9 displays the average categorization results. The model that optimally explained the data consisted of a single effect for the step factor, $b = -0.941$, $p < .001$. If the precursor factor was included, it did not show a significant effect, $b = 0.051$, $p = .452$.

Experiment 4b: Low-amplitude parts The results are shown in the bottom left panel of Fig. 9. The optimal model consisted of an effect for the intercept, $b = -0.392$, $p = .046$, which indicates that the probability of a nonspeech /pɪt/ response was smaller than .5, and an effect for step, $b = -1.166$, $p < .001$. If the precursor factor was included, it did not result in a significant effect, $b = -0.062$, $p = .394$.

Experiment 4c: All but spectral rotation (200-Hz F1 range) The top right panel of Fig. 9 displays the average categorization results. The optimal model showed a main effect for the factors step, $b = -2.110$, $p < .001$, block, $b = -0.068$, $p = .002$, and precursor, $b = 0.247$, $p = .010$. The precursor effect reflects a small but significant effect in the predicted direction. Additionally, a three-way interaction between the factors step, precursor, and block was found, $b = 0.070$, $p = .003$. This reflected the fact that in the first half of the experiment, the effect of precursor was more pronounced on one side of the continuum, whereas it was more pronounced on the other end of the continuum in the second half of the experiment.

Experiment 4d: All but spectral rotation (60-Hz F1 range) The bottom right panel of Fig. 9 displays the results. The optimal model showed a main effect for the factors step, $b = -1.050$, $p < .001$, and precursor, $b = 0.137$, $p = .0498$. The latter reflects a very small but significant effect in the

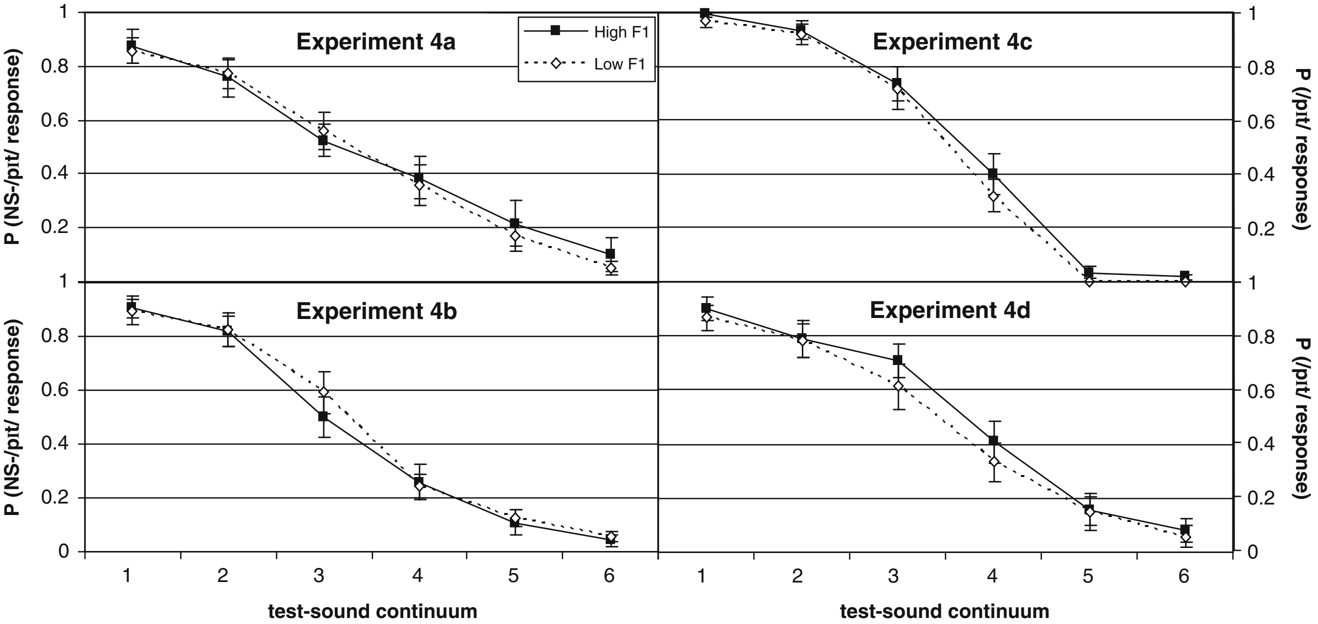


Fig. 9 Experiment 4. Upper left panel: Results for Experiment 4a. The graph displays probabilities of "nonspeech /pɪt/" (NS-/pɪt/) responses to the continuum of nonspeech targets ranging from "nonspeech /pɪt/" (step 1) to "nonspeech /pɛt/" (step 6). Targets were presented after nonspeech precursors that were created from both the high-$F1$ and low-$F1$ nonspeech sounds in Experiment 2, which now did have pitch movement. Bottom left panel: Results for Experiment 4b. The graph displays the same probabilities as those in the top left panel. Targets were presented after nonspeech precursors that were created from both the high-$F1$ and low-$F1$ nonspeech sounds in Experiment 2, which now had low-amplitude parts. Upper right panel: Results for Experiment 4c. The graph displays probabilities of /pɪt/ responses to the continuum of targets ranging from /pɪt/ (step 1) to /pɛt/ (step 6). Targets were presented after precursors that were created from both the high-$F1$ and low-$F1$ precursors in Experiment 2, which now had been spectrally rotated back. The targets' $F1$ frequencies ranged over 200 Hz. The result of this was that the formant structures appeared in their original positions. Error bars reflect standard errors. Bottom right panel: Results for Experiment 4d. The graph displays the same probabilities as those in the top right panel. The precursor was the same as in Experiment 4c. The targets were the same as those used in Experiment 1a and, thus, covered an $F1$ range of 60 Hz. Error bars reflect standard errors

predicted direction. A two-way interaction was found between the factors step and block, b = -0.025, p = .030, indicating that responses became more categorical as the experiment progressed.

Additionally, a comparison was made between the effects in Experiment 1a and the effects in Experiment 4d, including the experiment factor (levels: Experiment 1a [-1] vs. Experiment 4d [1]). The optimal model showed main effects for the factors step, b = -0.908, p < .001, and precursor, b = 0.338, p < .001. Interactions were found between the factors experiment and step, b = -0.140, p < .001, and between experiment and precursor, b = -0.202, p < .001. The latter shows that the effect of precursor was strongly reduced in Experiment 4d (when the precursors lacked speechlike prosodic characteristics). A three-way interaction was found between the factors experiment, step, and block, b = -0.026, p < .001, which reflects that, in Experiment 1a, participants' responses became increasingly less categorical, whereas in Experiment 4d, responses became increasingly categorical as the experiment progressed.

Discussion

Experiments 4a and 4b show that reintroducing pitch movement or silent periods in the signal does not change the influence of a precursor sentence on subsequent targets. Spectrally rotating the materials of Experiment 2 again (such that the formant structures appeared in their original locations and the signal had a speechlike spectral tilt) resulted in small but significant normalization effects in Experiments 4c and 4d. Note that Experiment 1b and Experiment 4c are, in a way, opposites. The manipulation that was applied to the stimuli in Experiment 1b (spectral rotation) was the only manipulation that was not applied to the materials in Experiment 4c. It thus appears that a signal with a speechlike spectral tilt and formant structures can induce a normalization effect but that this is not the only signal characteristic that can do so. The combined speech-like acoustic aspects that were available in the materials in Experiment 1b led to a numerically larger effect than did the speech-appropriate formants in Experiments 4c and 4d.

The comparison between Experiments 1a and 4d showed that the influence of a precursor was attenuated when the precursor lacked a number of speechlike prosodic characteristics. It thus appears that the speechlike acoustic aspects that were available in the materials in Experiment 1b and the speech-appropriate formants in the materials in Experiment 4c and 4d induced similar compensation processes. In speech materials like those in Experiment 1a, all these speechlike acoustic aspects are available. This would explain the relatively larger normalization effect found in Experiment 1a.

Experiment 5

An alternative explanation of the occurrence (or absence) of normalization effects in the different experiments presented so far is that such differences are the result of differences in how speechlike the stimuli were perceived to be. The stimuli in Experiments 1b, 4c, and 4d (where normalization was found) not only were acoustically similar to speech, but also could have been perceptually more similar to speech, while those in Experiments 2, 4a, and 4b (where no normalization was found) may have been both acoustically and perceptually less similar to speech. This would suggest that a listener's overt perception of the "speechiness" of the materials could induce a compensatory strategy, resulting in normalization effects only for those materials that listeners judge to be similar to speech. Although informal discussion with the participants after the experiments does not support this interpretation, it is a possibility that deserves more direct investigation. Experiment 5 was set up to investigate this matter by asking a new group of participants to rate the different precursor signals that were used in this series of experiments on their similarity to speech. If perceived "speechiness" influences the amount of normalization, it would be expected that the precursors in Experiments 1b, 4c, and 4d (which elicited normalization effects) would be judged to sound more speechlike than the materials in Experiments 2, 4a, and 4b. A lack of these patterns would argue against a role of "speechiness" in the induction of normalization effects.

Method

Participants

Sixteen further participants from the Max Planck Institute for Psycholinguistics participant pool were recruited and tested. They received a monetary reward for their participation.

Materials

Eight types of precursors were presented. Five consisted of the same precursors that were used for Experiments 1b, 2, 4a, 4b, and both 4c and 4d. These five types were each presented in their low-$F1$, neutral (i.e., as used in the third part of the training phases of the experiments), and high-$F1$ versions, resulting in 15 different precursors. Additionally, participants also rated three more types of sounds: noise, unmanipulated speech, and band-pass filtered speech. The noise precursors consisted of three versions: noise with the same amplitude envelope as the precursors from Experiment 1, noise with the same long-term average spectrum as the neutral precursor from Experiment 1, and a combination of these two manipulations. The unmanipulated speech consisted of three sentences, spoken by the same speaker and recorded during the same recording session as the materials for the previous experiments. The band-pass filtered speech consisted of filtered versions (200–2500 Hz, as for the materials in the previous experiments) of these unmanipulated speech sentences. This resulted in a total of 24 (5 × 3 for the precursors from the previous experiments and 3 × 3 for the additional types of carriers) different precursors. The precursors were randomly presented 3 times, during consecutive blocks. The unmanipulated speech and the noise conditions provided perceptual anchors for the other conditions.

Participants were asked to indicate how speechlike they thought the stimuli were. Participants heard a precursor once and then made their judgment by moving a vertical cursor along a (51-step) horizontal bar with, on the left of the bar, "niet-spraak" (nonspeech) and, on the right, "spraak" (speech). Participants moved the cursor using the mouse wheel, which started in the middle of the bar on every trial.

Results and discussion

Figure 10 displays the average "speechiness" ratings. The precursors that were most extremely manipulated (those used in Experiment 2) were not rated as more or less speechlike than those in Experiments 4a (nonspeech with pitch movement, b = -0.618, p = .429) and 4b (nonspeech with low-amplitude parts, b = -0.222, p = .775), but they were rated as less speechlike than those in Experiments 4c and 4d (all nonspeech manipulations, except for spectral rotation, b = 10.70, p < .001). The precursors in Experiment 2 were rated as slightly more speechlike than the precursors in Experiment 1b (nonspeech with only spectral rotation, b = 1.736, p < .043). The latter result suggests that there is no relation between perceived “speechiness” and amount of normalization, since such an account would predict that the precursors used in the experiment with the most extremely manipulated stimuli (Experiment 2) would be judged considerably less like speech than the precursors used in the nonspeech experiment with only spectral rotation (Experiment 1b), because only the latter elicited normalization effects.

Furthermore, the precursors in the experiments with all but spectral rotation (Experiments 4c and 4d) were rated as far more speechlike than the precursors in Experiment 1b, b = 12.44, p < .001. This was found despite the fact that the normalization effects in Experiments 1b and 4d were numerically similar, with a smaller b value for the context effect in Experiment 4d than in Experiment 1b. These results provide further evidence against an account that suggests that normalization depends on the perceived “speechiness” of the precursors.

General discussion

It was tested whether listeners take a precursor signal into account when categorizing speech or nonspeech targets. Normalization was found—that is, an influence of a precursor on the perception of a target—with both speech and nonspeech sounds. For both speech and nonspeech targets, normalization varied with the exact acoustic shape of the precursor, however. If the precursor did not contain speechlike prosody—that is, amplitude and $f0$ variations—normalization was severely reduced for speech sounds (Experiments 4c and 4d, as compared with Experiment 1a), and normalization disappeared for nonspeech sounds (Experiment 2, as compared with Experiment 1b). A final experiment tested whether the amount of normalization varied with the perceived "speechiness" of the precursors. This was not the case. Spectrally rotated stimuli with prosodic variation but an atypical spectral tilt created normalization effects similar to those for stimuli with a speechlike spectral tilt, but

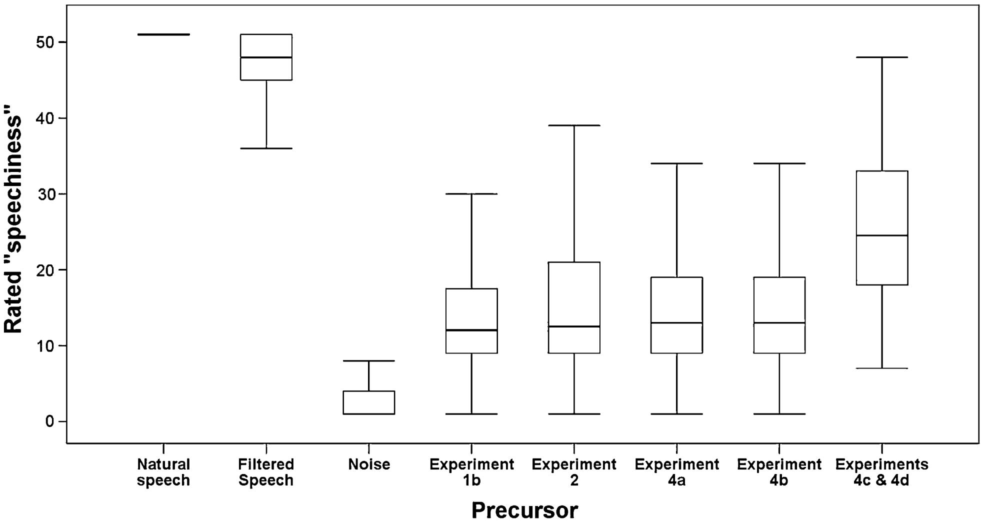


Fig. 10 Experiment 5: Boxplots representing "speechiness" ratings on a 51-step scale (0 = nonspeech, 51 = speech) for natural speech, band-pass filtered speech, noise, and the precursors used in Experiments 1b, 2, 4a, 4b, and both 4c and 4d

without prosodic variation. Nevertheless, the former were rated as much less speechlike than the latter.

One point of departure for the present study was the finding by Watkins (1991) and Watkins and Makin (1994) that normalization at central processing levels is restricted to stimuli containing spectrotemporal variation. Our results reveal additional restrictions, because only two out of the five manipulated stimulus sets induced normalization (see Table 1). One of these was a spectrally rotated version of the otherwise intact speech stimuli. These signals therefore contained speechlike acoustic aspects (prosodic aspects; see Table 1 for details). This result also shows that it is not necessary for precursor signals to consist of natural or interpretable speech for normalization to occur. Once stimuli contained speechlike acoustic aspects, central normalization processes did take place, probably in the form of compensation for LTAS. It thus seems that acoustic similarity to speech is a prerequisite for the LTAS of the precursors to influence the perception of subsequent targets. Interestingly, this prerequisite also applied if the materials contained manipulated speech materials, as in Experiments 4c and 4d. These precursor signals contained identifiable steady-state vowels but lacked prosodic variation.

To account for the present set of results, we suggest that, in compensation for LTAS at central processing levels, learning plays an important role. When listeners are confronted with different speakers in daily life, they may learn that the acoustic properties of speakers remain relatively stable. Given extensive exposure, listeners could therefore learn that taking the LTAS into account is beneficial while listening to speech. A striking parallel can be drawn with findings by Johnson, Strand, and D'Imperio, (1999), who reported an auditory identification shift for a *hood–hud* continuum for listeners who merely imagined listening to a male or a female speaker. In this situation, normalization is thus possible even without any acoustic precursor input. A learning approach can account for these findings, since it suggests that, through experience, listeners acquire the ability to normalize auditory input for speaker characteristics, be it LTAS or learned gender characteristics. A learning account can also accommodate the finding that the amount of central normalization varies with the exact nature of the precursor. Apparently, precursors with little prosodic variation are judged as less relevant for upcoming information, and therefore, their influence on the perception of the target signal is diminished or even obliterated.

An alternative to a learning account might be an enhanced auditory account in which the auditory properties directly determine (i.e., without learning) whether compensation for LTAS will take place. Listeners are sensitive to acoustic change, rather than to constancy (Kluender, Coady, & Kiefte, 2003; Kluender & Kiefte, 2006). Perceptual properties of static information are reduced in order to increase sensitivity to more informative information. This explains several findings of contextual influences, such as those by Watkins and Makin (1994, 1996) and Kiefte and Kluender (2008). From this viewpoint, it could be that the lack of normalization effects that was observed by Watkins (1991) with filtered noise precursors was the result of a reduction in the perceptual effect of the precursors as a result of their constant nature. However, the materials in Experiments 4d and 2, for example, did contain considerable spectrotemporal variation, such as the movement of $F1$ and $F2$ (or its nonspeech counterpart in the nonspeech experiments), but did not give rise to a normalization effect (Experiment 2) or gave rise only to a strongly reduced one (Experiment 4d). There were also considerable spectral differences between the precursors and the targets. Furthermore, the fact that listeners reliably categorized the stimuli in all the experiments shows that our precursor sets did not lead to a general lack of sensitivity to the $F1$ properties of our stimuli (or the properties of the spectrally rotated counterpart of $F1$).

The present set of experiments thus show that an alternative proposal, not dependent on learning, would have to go beyond the prerequisite of spectrotemporal variation that was proposed by Watkins (1991) and Watkins and Makin (1996). Such a proposal would have to account for the lack of normalization effects found in Experiments 2, 4a, and 4b and for the reduction of the effect in Experiment 4d, as compared with Experiment 1a. Other potential explanations of why normalization effects were not observed here might focus on fine-grained spectrotemporal properties and/or spectral tilt of the precursors. For instance, the rising spectral tilt in some of our nonspeech materials is unlike most sounds in nature and might induce perceptual effects unlike those found with natural sounds. In analogy to the approach that was taken in Experiment 4, such factors could be investigated by analyzing the effect of individual acoustic aspects of the signal. At this point, however, an obvious single acoustic candidate property fails to emerge from the data.

Although more research is needed on a learning-based account, it nevertheless seems to be the simplest way to extend a general auditory theory so as to also explain the present findings. Importantly, however, we do not want to suggest that all perceptual normalization is a consequence of learning. Instead, the total amount of normalization that is found in speech perception is likely to be a combination of normalization/compensation processes that take place at different stages of processing (Holt & Lotto, 2002; Mitterer, in press; Watkins, 1991). We attempted to focus on normalization that takes place at higher and more central levels in the processing stream by introducing a 500-ms precursor–target interval. The importance of this manipulation becomes clear from findings of normalization effects obtained with speech-shaped signal-correlated noise sounds

(Watkins, 1991; Watkins & Makin, 1994). Watkins found a large categorization shift at a 0-ms precursor–target interval when such precursor sounds were presented ipsilaterally but a complete absence of compensation when the precursor signals were presented contralateral to the target sounds (indicating that compensation took place at a peripheral level of processing). When these precursors were presented binaurally with a precursor–target interval of 160 ms, there was still a shift of about half the size of the shift that was found with ipsilateral presentation. As Watkins suggested, it is thus likely that at 160 ms, there is still some compensation due to peripheral mechanisms. This peripheral effect might be of the same type as that which causes the auditory afterimages found by Wilson (1970). Central compensation mechanisms add to the normalization that is the result of those earlier, more peripheral mechanisms.

A possible intermediate stage is formed by perceptual contrast effects (Holt & Lotto, 2002). These effects (also described as compensation for coarticulation effects) are a compensation that also gives rise to a categorization shift similar to that found with extrinsic normalization. These effects have usually been investigated by categorization of consonant pairs, such as /ga/ versus /da/, and show that listeners more often interpret an ambiguous sound from a /ga–da/ continuum as /ga/ when presented after /al/ than when presented after /ar/. Like extrinsic normalization, this shift can be explained as arising from contrastive spectral characteristics in the preceding syllable. A preceding syllable with more energy in higher frequency regions leads to the perception the low-frequency target, and vice versa for a preceding syllable with more energy in lower frequency regions. Holt and Lotto (2002) and Lotto, Sullivan, and Holt (2003) investigated the time course over which local contrast effects generally decay and found that they were restricted to durations no longer than 400 ms, which makes it unlikely that these effects had a large contribution in the effects reported in this article. Importantly, however, Holt and Lotto showed that these effects are also obtained with contralateral presentation. They thus argued that these effects are different from peripheral adaptation effects.

These local contrast effects do appear to operate in a general auditory way. Lotto, Kluender, and Holt (1997) reported demonstrations of local contrast effects in birds. Furthermore, Holt (2006) found compensation effects with notched-noise precursor stimuli (although peripheral effects were not controlled for, since the stimuli were presented bilaterally and with a 50-ms interval). Either of these two relatively early levels could also be the level of operation for the mechanism causing effects of speech and musical precursors on immediately following musical targets (Stilp et al., 2010). Interestingly, Stilp et al. also addressed the issue of learning with these materials. They reported that participants' musical experience did not influence the size of the compensation effect that was found. These combined findings suggest that effects at these early levels of processing are indeed of a general nature, indifferent to the exact acoustic nature of the preceding signal and not dependent on learning.

It appears, however, that these two relatively early compensation mechanisms do not influence categorization with precursor–target intervals of 500 ms or more. If extrinsic normalization were due to an early compensation mechanism that is indifferent to the fine-grained spectral and temporal properties of the signal but can apply over at least 500 ms, normalization ought to have been found in Experiments 2, 4a, and 4b (or, for that matter, with the signal-correlated noise materials used in Watkins, 1991). The materials in all these cases had the LTAS properties to induce compensation, yet no effect was found.

The present results, however, also demonstrate that the compensation effects of a possible learning mechanism are not restricted to speech per se. Acoustic signals that are, at a gross level, sufficiently similar to speech signals are subject to similar types of normalization. The fact that gross acoustic similarity is enough to induce normalization effects is in accordance with normalization effects that have been found with temporally reversed precursor syllables (Watkins, 1991; Watkins & Makin, 1994), rather than with normal sentences. An important additional implication of a learning approach is that learning to adjust perception for acoustic context signals is indeed by no means restricted to the processing of speech sounds. Any sound structure that shows LTAS constancy over time (and for which it would be beneficial to normalize, e.g., due to overlapping sound categories) could, in principle, evoke learned normalization processes.

In sum, we suggest that normalization for context at short ISIs is driven by automatic auditory processing that is independent of learning, while normalization at longer ISIs is influenced by learning. This suggestion seems to be at odds with one particular set of findings, however. Holt (2005) reported an influence of a long sequence of steady sine wave tones on categorization of a subsequent ga–da continuum, despite an ISI of up to 1.3 s. There are two possible ways to interpret this discrepancy. A first explanation could be that the context effects found by Holt (2005) could be of a different type: Such effects have generally been investigated with changes in categorization of (transient) consonants, whereas our investigation focused on shifts in identification for (more stationary) vowels. This is in line with a suggestion by Mitterer (2006), who investigated compensation for coarticulatory lip-rounding in fricative perception and found a pattern of results that was completely opposite to the pattern found for stops. Just as in the present case, this suggests that normalization for transient stimuli (stops) may be different from normalization for stationary stimuli (vowels and fricatives). Alternatively, Holt

(2005) argued that the persistence of the effects obtained with these acoustic histories, as compared with the rapidly diminishing effects obtained with temporally adjacent contexts (Holt & Lotto, 2002; Lotto et al., 2003), supports the idea that contrast effects may exist at multiple time scales. This supports the idea that local contrast effects and normalization effects over longer time scales are due to functionally different processing levels. This makes it possible that Holt (2005) reported on effects that should be attributed to the same processing level as the effects reported here. The acoustic context in the study of Holt (2005) contained an alteration of tones and silences, which may be sufficient prosodic variation for learned normalization to be engaged.

Another apparent contradiction with our suggested role of familiarity stems from a finding by Remez, Rubin, Nygaard, and Howell (1987), who reported extrinsic normalization effects (also at an ISI of 500 ms) with sine wave replicas of precursor sentences and targets similar to those used by Ladefoged and Broadbent (1957). On the one hand, such sine wave models lack the acoustic complexity of speech signals, and it is unlikely that listeners will have been exposed to these kinds of materials very often. On the other hand, sine wave models can be interpreted as speech and, therefore, appear to share some crucial auditory and prosodic aspects with speech as well. In addition to the availability of formantlike structures, sine wave models have specific time-varying phonetic characteristics, such as low-amplitude parts and syllable amplitude onsets and offsets (attack and decay structures), that are typical of speech sounds. As such, sinewave replicas of speech sounds could be sufficiently similar to familiar speech sounds to engage a learned normalization process.

To summarize, the present set of results suggests that extrinsic normalization at central processing levels is not exclusively the result of an all-purpose auditory compensation mechanism that is indifferent to the exact nature of those stimuli. The nonspeech precursors that were spectrally complex but had relatively little similarity to speech did not produce normalization effects on their nonspeech targets. Importantly, however, normalization was not restricted entirely to speech stimuli: Once nonspeech stimuli were substantially similar to speech materials on an acoustic level, normalization effects were found. We suggest that the simplest and most parsimonious way to explain these results is to assume that an important component of the perceptual normalization of vowels over longer time spans could, therefore, be an auditory mechanism that has been acquired over a lifetime of experience with different speakers or acoustic events. This perceptual mechanism could be a learned response to the covariations in natural sound patterns. Learning is a useful way to deal with the contextual influences that have not been resolved at lower levels in the processing stream. Lower level compensation mechanisms deal with covariations at shorter latencies and can do so because, at short latencies, it is more likely that a new sound originates from the same source. At longer latencies, however, stability of a sound source is less likely, and therefore, general mechanisms can be potentially harmful if they compensate perception of a new sound using the wrong source characteristics. A learning-based mechanism is a relatively simple way to deal with more situation-dependent covariations at longer latencies, because such a mechanism can adjust to specific situational demands. Longer term extrinsic normalization appears to operate as one of an array of other contextual mechanisms, such as lower level compensation mechanisms (Wilson, 1970), intrinsic normalization (Nearey, 1989), audiovisual integration (Massaro & Jesse, 2007), and lexically guided retuning of speech perception (Norris, McQueen, & Cutler, 2003). These mechanisms act in concert to resolve variability in speech signals.

Acknowledgements This work is part of the PhD work of the first author, funded by the Max Planck Gesellschaft. Part of it was presented as a poster at the 156th meeting of the Acoustical Society of America. The authors are grateful for comments on the manuscript by Alexandra Jesse, Michael Kiefte, and an anonymous reviewer.

Appendix 1

Pretest

This pretest was designed to establish the step size needed in order for listeners to reliably distinguish the /pɪt/ and /pɛt/ stimuli and the nonspeech versions of these stimuli, using a staircase procedure.

Method

Eight Dutch participants from the Max Planck Institute participant pool were recruited. They received a monetary reward for their participation. Sounds were created ranging from [pɛt] to [pɪt] (using the base sound and construction method described in Experiment 1). These sounds were combined in pairs. The first level consisted of [pɛt] (0-Hz *F*1 decrease) and [pɪt] (200-Hz *F*1 decrease, relative to [pɛt]). This level will now be referred to as the "0–200" pair. The other levels consisted of increasingly smaller differences (e.g., "10–190," "20–180"). The difference decreased in steps of 20 Hz until it was 100 Hz. It then decreased in steps of 10 Hz until it was 50 Hz. Then it decreased in steps of 4 Hz until it was 22 Hz, and finally, it decreased in steps of 2 Hz until it was 0 Hz. There were 28 steps in total. For catch trials, one of the sounds from each level appeared twice (e.g., 87–87). To create the nonspeech stimuli, all of these materials were spectrally rotated.

Participants were instructed that the experiment was designed to establish the limits of their hearing. They were told that, on every trial, they would hear two sounds that were either the same or different, and they had to respond by pressing response keys labeled "Hetzelfde" (*same*) or "Verschillend" (*different*). Participants judged both the speech version and the spectrally rotated version (order was randomized over participants). Each version started with a frequency difference of 200 Hz (the "0–200" pair).

When participants correctly responded to a block of stimuli (four pairs; two same, two different, in randomized order, from the same difficulty level), they would move to the next, more difficult level, until they reached their threshold (an upward run). If a participant responded incorrectly, the difficulty level would decrease by two levels, every time, until the participant completed a block correctly (a downward run). The increase in difficulty in an upward run involved large level changes before the first and second downward runs (five and three levels, respectively) and increases of only one level thereafter (up to a total of seven upward runs). A participant's discrimination level was calculated as the average lowest level at which that participant started his or her upward runs, on the basis of the last five upward runs.

Results

The results are displayed in Fig. 11, which displays the average lowest level at which an individual participant started an upward run. The difference levels to be used in the main experiments were selected to be those that all 8 participants could discriminate (i.e., the lowest levels at which all the participants showed, on average, discrimination). For the speech version, this level was set at a difference of 60 Hz. For the spectrally rotated version, this level was set at a difference of 200 Hz.

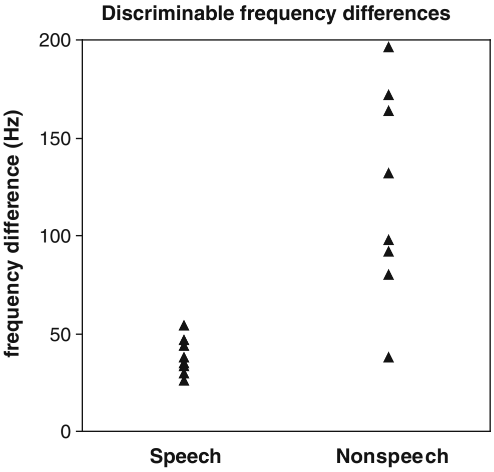


Fig. 11 Pretest: Participants made same–different judgments on pairs of stimuli in a staircase procedure. Triangles represent individual participants and show the average levels at which those individuals reliably discriminated the stimuli

Appendix 2

Table 2 Number of blocks needed to reach criterion in the training phases of each experiment

Experiment	Training phase						Participants
	Part 1		Part 2		Part 3		
	Mean	Max	Mean	Max	Mean	Max	
1a	9	29	4.5	11	3.1	4	8/9
1b	6.8	16	4.6	9	6.4	20	8/10
2	6.1	25	3.4	6	3.3	4	8/8
4a	4.3	7	3.4	6	5.8	17	8/11
4b	4.5	12	3.3	4	6.9	18	8/8
4c	3.3	4	3	3	3	3	8/8
4d	6.5	12	4.3	10	5	13	8/10

Mean number of blocks (mean) and maximum number of blocks (max) across participants are shown for each part of the training phase. Also shown are the number of participants who reached criterion, out of the total number of participants who were tested

References

Adank, P., van Hout, R., & Smits, R. (2004). An acoustic description of the vowels of Northern and Southern Standard Dutch. *Journal of the Acoustical Society of America, 116*, 1729–1738. Available at http://dx.doi.org/10.1121/1.1779271

Bates, D. M., & Sarkar, D. (2007). lme4: Linear mixed-effects models using S4 classes, R package version 0.99875-6.

Boersma, P., & Weenink, D. (2005). Praat: Doing phonetics by computer [Computer software]. Retrieved from http://www.praat.org

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science, 16*, 305–312. doi:10.1111/j.0956-7976.2005.01532.x

Holt, L. L. (2006). The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization. *Journal of the Acoustical Society of America, 120*, 2801–2817. doi:10.1121/1.2354071

Holt, L. L., & Lotto, A. (2002). Behavioral examinations of the level of auditory processing of speech context effects. *Hearing Research, 167*, 156–169. doi:10.1016/S0378-5955(02)00383-0

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics, 27*, 359–384. doi:10.1006/jpho.1999.0100

Kiefte, M., & Kluender, K. R. (2005). The relative importance of spectral tilt in monophthongs and diphthongs. *Journal of the Acoustical Society of America, 17*, 1395–1404. doi:10.1121/1.1861158

Kiefte, M., & Kluender, K. R. (2008). Absorption of reliable spectral characteristics in auditory perception. *Journal of the Acoustical Society of America, 123*, 366–376. doi:10.1121/1.2804951

Kluender, K. R., Coady, J. A., & Kiefte, M. J. (2003). Sensitivity to change in perception of speech. *Speech Communication, 41*, 59–69. doi:10.1016/S0167-6393(02)00093-6

Kluender, K. R., & Kiefte, M. (2006). Speech perception within a biologically realistic information-theoretic framework. In M. A. Gernsbacher & M. Traxler (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 153–199). London: Elsevier.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29*, 98–104. doi:10.1121/1.1908694

Lotto, A. J., Kluender, K. R., & Holt, L. L. (1997). Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*). *Journal of the Acoustical Society of America, 102*, 1134–1140. doi:10.1121/1.419865

Lotto, A. J., Sullivan, S., & Holt, L. L. (2003). Central locus for nonspeech context effects on phonetic identification. *Journal of the Acoustical Society of America, 113*, 53–56. doi:10.1121/1.1527959

Luce, R. D. (1986). *Response times*. New York: Oxford University Press.

Massaro, D. W., & Jesse, A. (2007). Audiovisual speech perception and word recognition. In M. G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 19–35). Oxford: Oxford University Press.

Mitterer, H. (2006). On the causes of compensation for coarticulation: Evidence for phonological mediation. *Perception & Psychophysics, 68*, 1227–1240.

Mitterer, H. (in press). Recognizing reduced forms: Different processing mechanisms for similar reductions. *Journal of Phonetics.* doi:10.1016/j.wocn.2010.11.009

Moore, B. C. J. (2003). *An Introduction to the psychology of hearing* (5th ed.). San Diego, CA: Academic Press.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85*, 2088–2113. Retrieved from http://www.utdallas.edu/~assmann/aud6306/nearey89.pdf

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*, 204–238. doi:10.1016/S0010-0285(03)00006-9

R Development Core Team (2008). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

Remez, R., Rubin, P., Nygaard, L., & Howell, W. (1987). Perceptual normalization of vowels produced by sinusoidal voices. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 40–61. doi:10.1037/0096-1523.13.1.40

Repp, B., & Liberman, A. M. (1987). Phonetic category boundaries are flexible. In S. N. Harnad (Ed.), *Categorical perception* (pp. 89–112). New York: Cambridge University Press.

Stilp, C. E., Alexander, J. M., Kiefte, M., & Kluender, K. R. (2010). Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets. *Attention, Perception, & Psychophysics, 72*, 470–480. doi:10.3758/APP.72.2.470

Watkins, A. J. (1991). Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America, 90*, 2942–2955. doi:10.1121/1.401769

Watkins, A. J., & Makin, S. J. (1994). Perceptual compensation for speaker differences and for spectral-envelope distortion. *Journal of the Acoustical Society of America, 96*, 1263–1282. doi:10.1121/1.410275

Watkins, A. J., & Makin, S. J. (1996). Effects of spectral contrast on perceptual compensation for spectral-envelope distortion. *Journal of the Acoustical Society of America, 99*, 3749–3757. doi:10.1121/1.414981

Wilson, J. P. (1970). An auditory after-image. In R. Plomp & G. F. Smoorenburg (Eds.), *Frequency analysis and periodicity detection in hearing* (pp. 308–318). Leiden, The Netherlands: Sijthoff.