The Influence of Acoustically Manipulated Similarity on Exemplar Effects

MA Thesis Language and Communication, Radboud University.

Tim Zee
Student number: 4647149

September 2017

Supervisor: Annika Nijveld, MA
Second reader: Prof. Dr. Mirjam Ernestus

Abstract

Auditory identity priming experiments generally show that participants respond more quickly and accurately to stimuli which are repeated with similar phonetic characteristics. However, these exemplar experiments usually do not control which phonetic features contribute to similarity between repetitions. Often this means that the degree of similarity between primes and targets is not tightly controlled. As a result, these studies generally provide limited insight into the relation between similarity and exemplar effects. The present research attempts to give a more detailed description of this relationship by manipulating a single perceptual parameter that has been shown to affect the similarity between prime and target stimuli: pitch.

In the first experiment of this study, participants adjusted the pitch differences between words through the resynthesis of fundamental frequency (f_0) in order to find the acoustic correlates of increasing pitch differences. In the second experiment, these pitch differences were used to create conditions of decreasing similarity between prime and target stimuli. It was hypothesized that a larger decrease in similarity between prime and target would lead to larger decrease in priming than a smaller decrease in similarity. However, the results showed that only the smaller pitch difference between prime-target pairs was associated with a decrease in priming effect. The increased pitch difference did not seem to cause any decrease in priming effect whatsoever, let alone a greater decrease in priming than the smaller pitch difference.

The inconsistent results of the exemplar experiment are not in line with previous research into the relation between similarity and exemplar priming. However, similarity is not the only factor that influences priming effects. It is possible that another factor in exemplar activation caused an increased priming effect that masked the inhibitory effect of the large pitch difference between the primes and targets.

Table of Contents

1. Introduction

In the past few decades, linguistics has seen a great deal of discussion about how listeners understand and produce words. It seems easy enough: when listeners hear a word they recognise the pronunciation of that word and link it to the meaning that is associated with those sounds. The problem with that explanation is that a pronunciation is a physical thing that can be measured directly, whereas a meaning or concept is a psychological thing, and at the current state of the art we do not really know yet how the physical sounds are linked to the concepts in our minds. The problem is made even more complicated by the fact that every pronunciation of a word is measurably different. So, somehow the psychological process that links lexical meaning to pronunciation must be able to deal with theoretically unlimited physical variation. This thesis takes a closer look at one of the theories that has been proposed to deal with this problem: Exemplar Theory.

Exemplar Theory (see Pierrehumbert, 2001 for an introduction) proposes that each word is stored with multiple representations, which are called exemplars. These representations reflect the different pronunciations that have been encountered for that word. In terms of word comprehension this means that whenever a listener encounters a word, the pronunciation of that word will be matched to the representations that resemble it (Ernestus, 2014, pp. 29-30). In other words, Exemplar Theory deals with variation by including it in the psychological representation of words. Some of the most convincing evidence in favour of this account comes from auditory identity priming experiments.

The results from word priming experiments have shown that increased similarity (e.g. the same voice) between the pronunciation of a previously heard word and a newly encountered pronunciation of the same word leads to faster and/or more accurate responses to the second pronunciation (Luce & McLennan, 2005, p. 599; e.g. Craik & Kirsner;). These so-called exemplar effects suggest that the first occurrence of the word creates an exemplar and that the similarity between that exemplar and the second occurrence of that word influences the speed and accuracy of word recognition (Goldinger, 1998, p. 254).

However, in most exemplar studies that use the priming paradigm, similarity is manipulated across multiple phonetic parameters. For instance, speaker voice variation is not only reflected in pitch, but also involves a particular voice-quality and speaker-specific articulations. Because it is unknown how the respective phonetic features contribute to similarity, it is hard to control to what extent any two versions of the same word are treated as similar during word comprehension. Consequently, many previous exemplar experiments have measured the differences between listeners' responses to identical repetitions and their responses to repetitions with some uncontrolled amount of variation. By manipulating a single perceptual dimension (*Pitch*), the present research hopes to investigate how similarity relates to exemplar effects in greater resolution.

Section 2 provides some more background information about the way similarity is conceptualised in exemplar models, reviews how previous priming studies have implemented similarity, and formulates specific research questions and hypotheses. Section 3 describes the pitch adjustment experiment we used to control the perceptual similarity between the first and second occurrences of words in the priming experiment that is discussed in Section 4. Finally, Section 5 summarizes the findings and discusses some limitations of this research that warrant further investigation.

2. Theoretical Framework

2.1 Similarity in Exemplar models

In order to make the concept of similarity more specific, this section describes a popular formal Exemplar model: The Generalised Context Model (GCM; Nosofsky, 1984). The following is by no means an exhaustive description of the model and is only meant to illustrate the nature of similarity in Exemplar theory. The GCM is a categorization model that can be applied to all kinds of stimuli. Applied to spoken word comprehension, this means that the categories correspond to words, and that each word category consists of exemplars.

Like many other exemplar models (e.g. Hintzman, 1986; Medin & Schaffer, 1978), the GCM holds that the perceivable physical properties of stimuli are stored as one or more psychological features. It is important to note that there is not necessarily a one-to-one mapping of physical and psychological features. Each stimulus, then, can be represented by a feature vector, with a certain value for each feature. Because a set number of features is assumed, different stimuli can be represented as points, or exemplars, in the same multidimensional feature space. In order to simulate actual perception experiments, such a space can be created by processing the mean responses to all stimuli using a Multi Dimensional Scaling technique (MDS; Kruskal & Wish, 1978).

Using this notion of an *M*-dimensional psychological space, the distance *d* between an incoming exemplar *i* and an existing exemplar *j* can be calculated according to the formula in 1 (adopted from Nosofsky & Johansen, 2000):

$$d_{ij} = \left[\sum_{m} w_m \cdot \left| x_{im} - x_{jm} \right|^r \right]^{1/r} \tag{1}$$

where x_{im} and x_{jm} represent the feature values for exemplar *i* and *j* on dimension m. The value of *r* depends on which distance metric is used, and can be ignored for the present purpose. The value of w_m represents the attention weight *w* of dimension m. As will be explained later on, this concept is important with regard to the relationship between perceptual similarity and exemplar effects. As expected, the similarity *s* between two exemplars i and j is inversely related to their distance *d*: similarity increases as distance decreases, see the formula in 2 (adopted from Nosofsky, 1991b):

$$s_{ij} = \exp\left(-d_{ij}^{p}\right) \tag{2}$$

Here, the value of *p* depends on the type of experiment that is modelled and is not important for the current theoretical discussion. Using this similarity measure, the GCM calculates the activation of exemplar *j* as the result of exemplar *i* according to equation 3 (adopted from Nosofsky, 1991a).

$$a_{ij} = M_j \cdot s_{ij} + e_j \tag{3}$$

In this formula, M_j represents the memory strength of exemplar *j* and e_j introduces random noise. The memory strength variable is especially interesting because it allows for factors unrelated to similarity to affect processing. Nosofsky suggests that the frequency and recency of exemplar *j* should play a role in defining memory strength (1991b, p. 135). Others have suggested that the salience of an exemplar also contributes to memory strength (Drager &

Kirtley, 2016, p. 13). In this context, salience is broadly defined as “the degree to which something stands out relative to other, neighboring items” (Drager & Kirtley, 2016, p. 12). According to Drager and Kirtley, this includes properties of the stimulus itself but also pertains to expectations based on earlier input (2016, p. 12). This notion can be related to the suggestion by Johnson that *top-down* information may influence exemplar activation (2005, p. 304). For instance, if previous input (i.e. preceding context) increases the likelihood of a particular word, the activation level of all exemplars of that word could be increased through this memory strength variable.

By looking at all activated exemplars, and summing the activation of exemplars that belong to the same category, the GCM calculates the probability that the incoming item is sorted into a category. This process is formalised in equation 4 for a situation with two categories *J* and *K*.

$$P(J|i) = \frac{\sum_{j \in J} a_{ij}}{\sum_{j \in J} a_{ij} + \sum_{k \in K} a_{ik}} \quad (4)$$

Using these probabilities, categorization tasks (e.g. sorting faces into learned categories; Nosofsky, 1991a) can be simulated, and the results of such a simulation can be compared to the categorizations made by actual participants. Furthermore, if it is assumed that higher (relative) activation levels lead to faster processing, response times (RTs) can be predicted as well (Goldinger, 1998, p. 254).

As a result of this formal description of an exemplar model, a number of important points arise, all of which relate to unknown variables. First of all, similarity is derived from the distance between two exemplars in a psychological multidimensional space. It is not known what these dimensions represent, or even how many there are. We only know that the input consists of the physical characteristics of the stimuli and that the output consists of a classification and a response time. Second, even if we did know the coordinates of two exemplars in the multidimensional space, it would not be straightforward to convert the distance between these coordinates into a similarity measure. Depending on the task at hand, listeners pay attention to different aspects of their environment. In exemplar models, this is formalised as the attention weights that highlight certain psychological dimensions and downplay others. Lastly, the relationship between similarity and exemplar activation is obscured by external variables such as salience that influence the memory strength of an exemplar.

The following section describes how previous research has attempted to control similarity and highlights the strengths and weaknesses of the different approaches.

2.2 Similarity in Previous Exemplar Experiments

This section deals with the different ways in which exemplar experiments have manipulated similarity and how these different approaches constrain the interpretation of the results. In light of that discussion, it is useful to start with a description of the basic exemplar priming paradigm. As this thesis is about an exemplar theoretic approach to spoken word comprehension, this section only considers word priming experiments.

In a repetition priming experiment, participants are repeatedly presented with the same item. The first presentation of an item is called the *prime* and its repetition is called a *target*. Experiments that apply this paradigm to a classification task (e.g. lexical decision, old/new judgements) have found that target stimuli are responded to more accurately and more quickly than primes (e.g. Forbach, Stanners & Hochhaus, 1974; Forster & Davis, 1984). This effect can of course be accounted for in an exemplar model through the memory strength variable (which is influenced by recency). However, an exemplar account also predicts that this priming effect should decrease as a function of decreasing similarity between prime and target. To test this prediction, the priming paradigm can be modified to include experimental conditions with different degrees of similarity between primes and targets: one portion of the target stimuli is chosen to be more similar to their respective primes than another portion of the target stimuli. This brings up an important question: How can similarity be manipulated?

Most exemplar experiments control similarity by creating two versions of the same word. For example, Craik and Kirsner (1974) created their stimuli by recording two speakers that read out the same word list. In their old/new judgement task, half of the words were repeated. For half of these repetitions, the same recording by the same speaker was used, resulting in an identical prime-target pair. For the other half, the recording of the same word by the other speaker was used, resulting in primes and targets that mismatched in speaker. This manipulation allowed the researchers to attribute the decreased performance in the *mismatching-voice* condition to a decrease in similarity between prime and target stimuli (Craik & Kirsner, 1974, p. 281). It is important to note that this type of manipulation provides no information about which psychological dimensions are primarily involved in the decreased similarity of the *mismatching-voice* condition. They could have been dimensions related to perceived gender, pitch, voice quality, or even dimensions that stem from differences that are completely unrelated to the identity of the speaker, such as random variation in the articulation of speech sounds. This lack of control over which psychological dimensions are manipulated makes it difficult to achieve a more granular view of how similarity relates to exemplar priming. For instance, it might be that only stimuli that are nearly identical facilitate word priming. In order to investigate this claim, it would be necessary to compare more than two conditions of decreasing similarity. However, without control over which dimensions are manipulated it is impossible to guarantee that one *mismatching-voice* condition is more different than another.

Another important consideration concerns the consequences of using identical primes and targets in the *matching-voice* condition of an experiment. As other researchers have noted (Hanique et al., 2013), using the same recording twice decreases the ecological validity of the experiment: In naturally occurring speech, even repetitions by the same speaker are not identical. For this reason, more recent exemplar experiments have used different recordings by the same speaker for prime-target pairs in the *matching-voice* condition (e.g. Hanique et al., 2013; Nijveld et al., 2015). However, when there is uncontrolled variation in both experimental conditions, a situation might arise in which the prime-target pairs in the *mismatching-voice* condition are just as similar or even more similar compared to those in the *matching-voice* condition. In fact, this lack of control might be the reason that one study with non-identical *matching-voice* primes and targets found exemplar effects that are speaker dependent (Nijveld et al., 2016). A similarity-based explanation for these results would be that some speakers are more consistent than others when they record multiple pronunciations of the same word.

The experiments described above illustrate how the extent to which the relation between similarity and exemplar effect can be investigated is limited by a lack of control over the psychological dimensions that make up similarity. One approach that has allowed researchers to have more control over this aspect is called Multi Dimensional Scaling (MDS). This approach takes *proximities* between different items as input and uses those to compute an optimal dimensional representation of the similarity between those items (Kruskal & Wish, 1978, p. 7). This is based on the assumption that those dimensions are somehow meaningful, be it physically, psychologically or otherwise. In order to find meaningful interpretations of the dimensions, regression analyses can be performed that use different features of the items as the dependent variables and the coordinates of the items along the dimensions as the independent variables (Kruskal & Wish, 1978, p. 36). In a series of exemplar experiments, Goldinger applied this technique to estimate the psychological dimensions involved in exemplar effects (1996). He presented participants with pairs of stimuli and asked them to indicate whether the pair consisted of two different words or a repetition of the same word. In total, 10 different speakers produced the stimuli, and in each pair, the items were pronounced by two different speakers. Based on the assumption that words repeated by similar speakers result in increased priming effects, Goldinger used the mean RTs for each speaker combination as proximities for an MDS analysis (1996, p. 1169). He found that a two-dimensional space defined by gender and pitch resulted in a good fit to the data (1996, p. 1169). He then used the Euclidean distances between different speakers in this 2-dimensional space as a similarity measure for subsequent exemplar priming experiments. The results showed that this similarity measure was associated with a significant improvement in both accuracy and latency in a number of different tasks (Goldinger, 1996). These results suggest that exemplar priming is not dependent on prime and target stimuli being identical. Rather, they provide evidence that the degree of word priming is indeed monotonically related to the similarity between target and prime, as would be expected if exemplar processing plays a role in spoken word comprehension. However, it should be noted that, due to the statistical nature of the method by which MDS defines its psychological dimensions, it should not be ruled out that the actual dimensions are somewhat different from the ones that were selected.

2.3 Similarity in the Present Research

As the previous section has detailed, a lack of control on the dimensionality of similarity has been the main obstacle in the way of a detailed description of the relation between similarity and exemplar priming. MDS addresses this problem by working backwards from differences in priming effects, using statistical techniques to isolate the psychological dimensions that best explain these differences. Although this technique has provided more insight into the role of similarity in exemplar processing, the psychological dimensions it identified and their relation to similarity should be verified by using other methodologies to control similarity.

2.3.1 General Methodology

The present research attempts to control similarity by limiting differences between stimuli to one of the psychological dimensions identified by MDS. However, as illustrated in Section 2.2, it is very hard to manipulate a single psychological dimension using natural phonetic variation. Fortunately, the phonetic exponent of pitch—fundamental frequency (F0)—can be manipulated acoustically using resynthesis techniques. Therefore, it is possible to control similarity through the perceptual dimension of pitch by estimating acoustic manipulations that result in increasing perceptual differences. One way to arrive at these estimates is by asking participants to rate the size of the perceptual differences between F0-manipulated stimuli. As we are interested in increased (non-identical) similarity, this rating task is best formatted as a comparison of two differences. Furthermore, this perceptual comparison between differences needs to be quantifiable in a way that is both intuitive to the participants and suitable for further use in exemplar experiments (i.e. the increase in perceptual similarity must be large enough to result in measurable effects). In light of these considerations, the present research uses a rating task in which listeners identify F0 manipulations that are perceptually twice as large as other F0 manipulations.

A number of existing experimental methods from psychophysical research can be adapted to find the F0 manipulations that correspond to a doubling in pitch difference. Two of these methods, the *method of constants* and the *method of limits*, make use of a forced choice paradigm. Applied to the present research, this approach would entail that participants indicate whether one F0 manipulation was perceptually more or less than double the size of another F0 manipulation. One of the F0 manipulations would then be increased or decreased and the same question would be posed. When this approach is combined with the amounts of words typically used in exemplar experiments, the required number of trials becomes so large that it may not be suitable for untrained participants. Fortunately, another psychophysical method exists which does not have this problem. The *method of adjustment* allows participants to directly control the physical manipulation of the stimuli. Applied to the present research, this means that, upon hearing a certain F0 difference, participants would create an F0 difference that is twice as large by directly controlling the F0 of the stimuli, listening to the result, and adjusting as necessary. By removing unnecessary iteration, this method requires much less time to converge on the participants' estimates of perceptual doubling in pitch. It should be noted that this increase in speed comes at the cost of experimental control (Ehrenstein & Ehrenstein, 1999, p. 1214), resulting in a greater potential for variability between participants and trials. However, as the present research is only interested in an estimate averaged over participants and items, this increase in variability can be compensated for by testing more participants.

Once the method of adjustment has identified an f0 difference that is perceptually twice as large as another f0 difference, these differences can be used to control similarity in an exemplar experiment. Most exemplar experiments manipulate similarity using two *within-participant* conditions: a *match* condition, in which prime and target are identical or similar, and a *mismatch* condition, in which prime-target pairs are less similar. In order to incorporate three levels of similarity, an additional within-participant level of similarity could be added. Applied to the current research, this design would entail three conditions of increasing pitch differences: a condition which consists of identical prime-target pairs, a condition made up of primes and targets with a small pitch difference, and a condition in which primes and targets differ by double the pitch difference of the previous condition. Alternatively, the additional level of similarity could be incorporated in a mixed design that consists of two between-subject conditions and two within-subject conditions. In this approach, each of the between-subject groups would contain a *match* condition with identical primes and targets and a *mismatch* condition with prime-target pairs that differ in pitch. In one of these between-subject groups, the shift in pitch between primes and targets in the *mismatch* condition would be twice as large compared to the other between-subject group. This design is summarized in Table 1.

Table 1: Exemplar priming experiment that controls similarity using two conditions of increasing pitch difference in a mixed design.

		Within-Subjects: Exemplar Priming	
		Match	**Mismatch**
Between-Subjects: Mismatch Difference	**Small**	Identical pitch	Small pitch difference
	Large	Identical pitch	Large pitch difference

A purely *within-participant* design is usually preferred in favour of a design with *between-participant* variables due to the increase in unsystematic variation associated with the latter design (Field, 2013, pp. 17-18). However, a combination of *within* and *between-participant* was used for the present research in order to prevent bias in potential categorisation processes that influence the perception of pitch. The bias that may arise in within-participant designs with more than two conditions is illustrated in Appendix A.

This section has described how acoustic manipulation of F0 can be used to create conditions of increasing perceptual similarity, but the nature of these F0 manipulations has not been discussed in much detail yet. The following section describes the most important considerations pertaining to the resynthesis of F0 in the context of exemplar experiments.

2.3.2 Pitch Manipulation

Although there is some precedent for the use of pitch resynthesis in exemplar experiments, its relevance for the present research is limited. Church and Schacter used *linear predictive coding* (LPC) algorithms to shift the F0 of stimuli by 10% (1994, p. 526) and found that participants performed significantly better on a word completion task if the word was primed in the same F0 (1994, p. 529). However, it has been shown that LPC manipulation introduces much more noticeable distortion than the more advanced *pitch-synchronous overlap-add* (PSOLA) technique (Moulines & Charpentier, 1990, p. 466) used in the present research. It follows that the similarity manipulation used by Church and Schacter (1994) may have included distortion in addition to a shift in F0. Consequently, a 10% shift in F0 should not necessarily be considered a sufficient manipulation for a reliable exemplar effect, especially as it did not result in exemplar effects in all of Church and Schacter's experiments (1994, p. 529). This raises the question of how large the shift in F0 can be before the distortion becomes unacceptable.

For PSOLA resynthesis, most research agrees that any manipulations between a 50 percent decrease and a 50 to 100 percent increase in F0 are acceptable (see review in Longster, 2003, p. 59). In other words, a stimulus with a natural F0 of 160 Hz can be manipulated to an acceptable standard within the range of 80 Hz to around 280 Hz. This range can be used to establish a guideline for the F0 difference that is to be perceptually doubled in the adjustment experiment (henceforth *Difference A*). Crucially, the size of *Difference A* must be chosen so that the F0 difference that corresponds to the perceptual double of *Difference A* falls within the range described above. Paradoxically, then, the range of suitable values of Difference A, which must be chosen beforehand, depends on the outcome of the adjustment experiment. Fortunately, previous research (Greenwood, 1990) has established a psychoacoustic scale (see equation 5) which can be used to give an indication of the outcome of the experiment[1].

$$ERB = 16.7 \log_{10}\left(1 + \frac{f}{165.4}\right) \tag{5}$$

This formula relates a psychophysical measure of pitch, *ERB*, to the physical measure of frequency, *f*. Applying this formula, we can calculate that a stimulus with a lower acceptability boundary of 80 Hz and an upper acceptability boundary of 280 Hz could be used to create a maximal perceptual difference of 4.32 ERB. Given the perceptual doubling task of the adjustment experiment, this means that the perceptual difference corresponding to *Difference*

[1] It should be noted that this scale is primarily derived from anatomical research (Greenwood, 1990) and as such may not account for influence of specific properties of the stimuli (such as vowel quality) on the perception of F0 differences. For that reason, it is not used as the primary method to control perception in this research.

A should be smaller than 2.16 ERB. According to the formula in (5), one way to create this perceptual difference would be through a physical difference between stimuli with an F0 of 80 Hz and 165 Hz respectively. However, the actual size of *Difference A* should be considerably smaller to account for the inconsistency between the predictions of the psychoacoustic scale and the perceptual results. Furthermore, the range of acceptable manipulation, which is a subjective measure to begin with, should also be established on the basis of experimentation with the materials that are used in the actual experiment.

Up until this point, the description of pitch manipulation has treated the F0 of a stimulus as a constant value. However, in natural word tokens, F0 changes over time due to a combination of physical pressures and the communicative functions of intonation. This is relevant for the present investigation as it has been shown that the same rate-of-change or slope of F0 results in different perceptions of intonation depending on the mean F0 (Hermes & Van Gestel, 1991; Nolan, 2003). In other words, simply raising or lowering all points on an F0 contour by the same physical amount could change the intonational perception of that F0 contour. In the context of the present research, this difference in intonational perception would be a confound for the perceptual difference in pitch. Most research on the influence of mean F0 on the perception of intonation has found that describing shifts in pitch on the ERB scale in (5) adjusts the F0 slope in a way that minimizes differences in intonational perception (Hermes & Van Gestel, 1991; Nolan, 2003). As such, this measure is used in the F0 manipulations of the adjustment experiment and henceforth F0 differences will be quantified using this measure.

2.4 Research Questions

Although previous experiments have provided evidence for some kind of relationship between psychological similarity and exemplar effects, only one methodology has found support for the more specific conceptualisation of similarity that is often assumed in exemplar research (see Section 2.1). More specifically, the application of MDS to the exemplar priming paradigm has resulted in the most convincing support for a multidimensional representation of spoken word comprehension by showing that a similarity measure between tokens derived from their position on perceptual dimensions is monotonically related to the priming effect between those tokens.

The current research attempts to verify those findings by conducting an exemplar experiment in which perceptual similarity is controlled along one of the dimensions identified by MDS using a different methodology. Specifically, the method of adjustment is applied to the perceptual dimension of pitch to find F0 differences of increasing size that can be used to create conditions of decreasing similarity (see Section 2.3). To that end, the following research question is answered:

RQ1 What is the size of the acoustic F0 adjustment that corresponds to a doubling of the perceptual difference between two stimuli?

Once the size of the acoustic manipulation that controls perceptual similarity has been estimated, the manipulation can be applied to an exemplar priming experiment to answer the main question of this thesis, see research question 2:

RQ2 Does an increase in the pitch difference between non-identical primes and targets lead to a decrease in priming effect?

It is expected that this decrease in priming effect will manifest itself in the response times to the target stimuli in the different conditions in Table 1. As in previous studies, an exemplar effect is expected in the form of longer response times in the *mismatch* condition and shorter response times in the *match* condition. Crucially, it is also expected that this latency differential will be greater if it involves a large (i.e. doubled) rather than a small pitch difference in the *mismatch* condition. A result that supports this hypothesis would, by extension, provide support for the way similarity is modelled in many exemplar theoretic accounts of spoken word comprehension.

In order to make sure that a potential decrease in priming effect can be attributed to the increase in pitch difference between primes and targets, two additional variables will be monitored. First of all, individual and between-subject group differences in participants' pitch sensitivity were gauged in a separate experiment. Although individual differences in perceptual acuity are unavoidable, between-subject group differences are not expected if all participants are recruited from the same population. Secondly, in order to keep track of any potential influences of the memory strength of previous representations (see Section 2.1), the participants' responses to the prime stimuli will also be measured. No significant differences in the response times to prime stimuli in the different similarity conditions are expected.

3. Pitch Adjustment Experiment

3.1 Methods

Procedure

For the present experiment the *method of adjustment* was adapted to find the F0 difference between two stimuli that was perceptually double the size of a smaller F0 difference. This was implemented as follows: Each trial involved three tokens of a word: version A, B, and C. The F0 of versions A and B could not be adjusted, and a fixed F0 difference corresponding to 1.2 ERB existed between them. The size of this difference was informed by the guidelines discussed in Section 2.3.2. The F0 of version C, on the other hand, was adjustable. A trial started with the auditory presentation of version A, a 500 ms silence, followed by version B. After a 500 ms pause, the trial continued with the second sequence, which consisted of version A and version C separated by another 500 ms silence. At any point during the trial, the participants could replay either of these sequences by clicking on their corresponding on-screen buttons. The button corresponding to the A-C sequence also functioned as a slider that could be dragged across a horizontal axis in order to change the F0 of version C. The participants were instructed to use this slider to make the perceptual difference between version A and C double the size of the difference between version A and B. Figure 1 shows a simplified version of the on-screen controls.

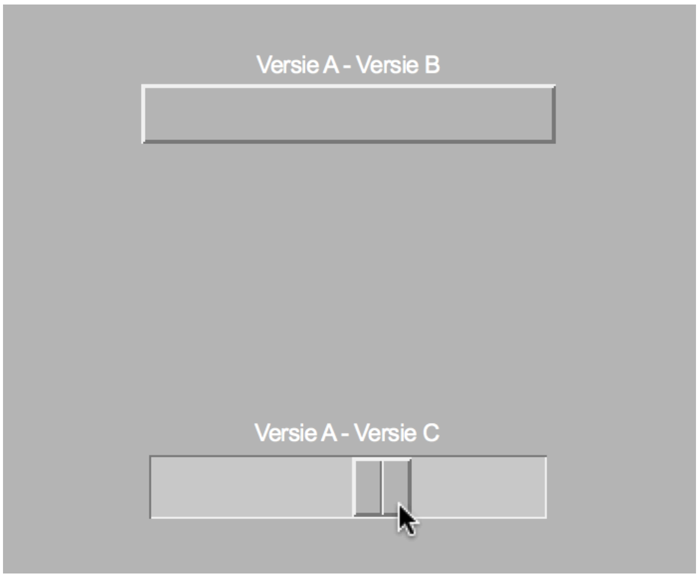

Figure 1: Simplified representation of the on-screen controls used by the participants in the pitch adjustment experiment. 'Versie' is the Dutch word for *version*.

In order to prevent participants from simply dragging the slider to the same position on each trial, the starting F0 of version C was varied across a range that was centred around the participants' expected F0 adjustment, which was derived from the formula in (5) and the results of pilot experiments. Combined with a slider that always started in the middle of its range of motion, this meant that in order to make somewhat constant pitch adjustments, the size and direction of the participants' visual adjustments would have to vary across trials. In other words, participants had to rely on auditory feedback rather than visual feedback to correctly carry out the task.

Materials

Recordings of 70 Dutch monosyllabic nouns were used to create the stimuli used in the adjustment experiment. An additional 5 words were used for practice trials. Only words with voiceless onsets and codas were used, because the PSOLA resynthesis of these consonants results in less distortion compared to voiced consonants (Longster, 2003, p. 161). Additionally, the nucleus of each word had to consist of one of the vowels or diphthongs in /a, o, e, ø, ɛi, œy, ɑu/. This restriction was implemented to limit the variation in the duration of voicing across different words. The list of candidate words was narrowed down further by omitting semantically salient (e.g. taboo) words and words that were very rare.

All words were read out loud multiple times by the same female speaker in a sound attenuated booth. Her speech was recorded through a head-mounted microphone and was digitized at a sample rate of 44100 kHz and a resolution of 16 bits per sample. For each word, the token with the least amount of voice creak was extracted and normalised to 70 dB. In preparation for PSOLA resynthesis, an automatic pitch analysis of each word was performed and manually adjusted if necessary. These analyses also revealed that the unmanipulated recordings had a mean F0 of 162.51 Hz with a standard deviation of 3.90 Hz.

Using the implementation of PSOLA in Praat (Boersma, 2001), a script was created that could automatically resynthesize stimuli at the F0 that was required by the program controlling the experimental procedure. At the start of each trial—and each time the participant would press the button corresponding to the difference between version A and B—the experimental software would instruct the Praat script that version A and B of a word needed to be resynthesized with a -0.4 ERB and a +0.8 ERB pitch shift respectively (resulting in a 1.2

ERB difference between version A and B). A similar but slightly more complex process applied to the presentation of the difference between version A and C. For the initial presentation of this difference, version A was again resynthesized with a -0.4 ERB shift, whereas the pitch shift of version C was word specific to prevent bias from visual cues. Each subsequent presentation of version C modified the initial word specific shift to reflect the adjustments to the slider. As a result of this adaptive implementation of the PSOLA algorithm, the resolution of the adjustments to the pitch of version C was not limited by pitch differences of premade stimuli.

Participants
Participants were recruited from the Max Planck Institute for Psycholinguingistics (MPI) participant pool and received € 8 for their cooperation. Two subjects were excluded from the main analysis (see Section 3.2 for motivation). The remaining 18 participants (12 female, 6 male) had a mean age of 21.1 years (SD = 2.1).

3.2 Results

First of all, the results were inspected for any abnormalities using individual scatterplots for each participant, as seen in Figure 2.

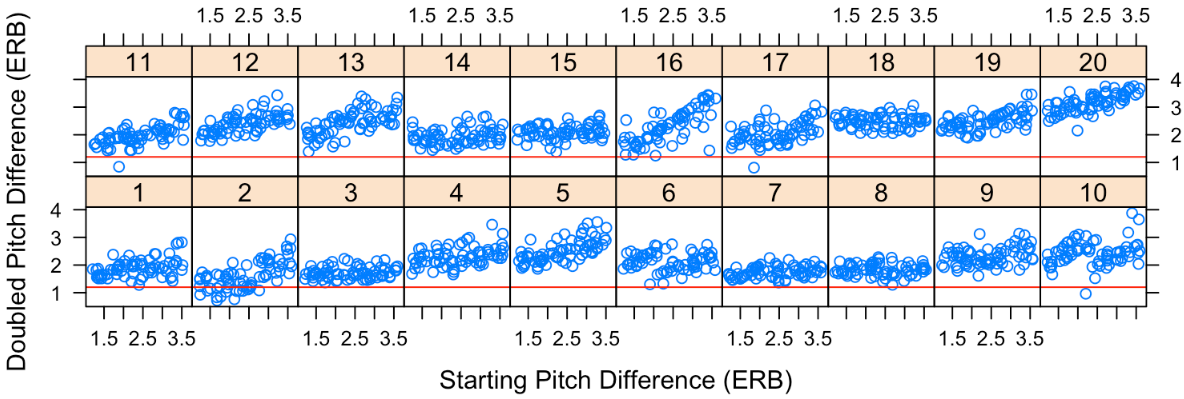

Figure 2: Individual scatterplots that plot the doubled pitch difference against the starting pitch difference.

The y-axes in Figure 2 plot each participant's estimates of what the pitch difference between A and C versions of each word should be. To put this into context, a red line representing the 1.2 ERB difference between versions A and B has been added to each scatterplot. As expected, almost all estimates of the perceptual double of a 1.2 ERB difference are greater than 1.2 ERB. However, 23% of participant 2's estimates are below the 1.2 ERB threshold, suggesting behaviour counter to the experimental task. For this reason, participant 2 was excluded from further analysis.

By plotting the starting difference between versions A and C of each trial on the x-axes, Figure 2 also illustrates that, for many participants, trials with a greater starting difference produced larger final estimates. This may reflect a tendency of the participants to not move the slider too far away from its starting position. This tendency would be problematic if participants made adjustments of the same size and in the same direction regardless of the starting differences. Luckily, as Figure 3 illustrates, it seems that this is not the case for most participants.

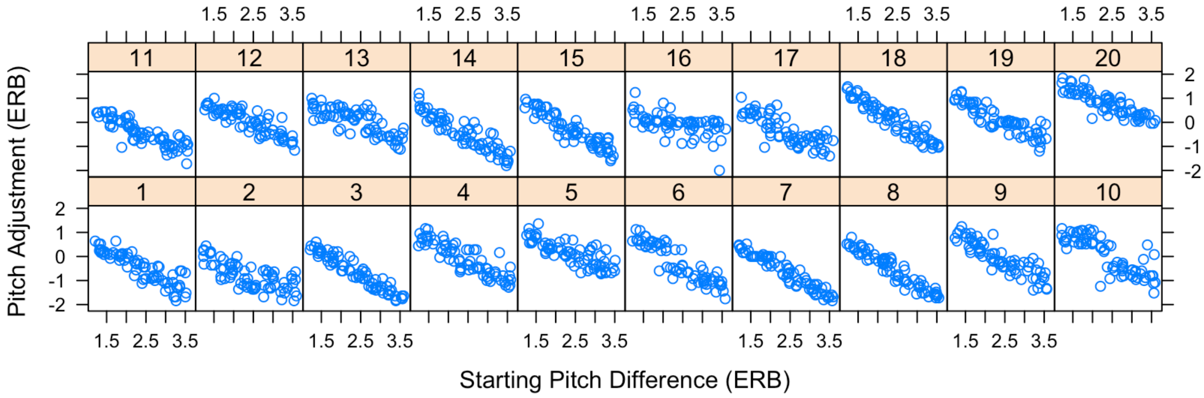

Figure 3: Individual scatterplots that visualize the relation between the starting pitch difference and the size of the adjustment made by the participants

However, there is one striking exception to the general negative relationship between pitch adjustment and starting difference: many of participant 16's adjustments deviate very little from 0, regardless of starting pitch. This suggests that participant 16 often paid little attention to the experimental task or was very insensitive to the comparison of the two pitch differences. For this reason, participant 16 was excluded from further analysis.

In order to arrive at an accurate mean estimate of the perceptual double of the pitch difference between the A and B versions of the stimuli, any remaining outliers were removed based on their fit to a model of the data. Figure 2 suggests a degree of between-participant variation regarding the mean estimate of the pitch doubling (compare participant 8 and 18) as well as the association between starting difference and pitch doubling estimate (compare participant 3 and 13). Therefore, a mixed effects model of doubled pitch difference was created with random intercepts for participants and words, by-participant random slopes for starting pitch difference, and a fixed effect of starting pitch difference. Adding fixed effects of participants' musical training, participants' strategies, the trial number, or the number of adjustments per trial did not result in significant improvements of the model. As such, Table 2 represents the final model.

Table 2: Linear mixed model of the participants' estimates of the doubled pitch difference.

	Model		***Refitted Model***	
Fixed effects	***B***	***t***	***B***	***t***
Intercept	1.79	22.47	1.82	22.45
Starting Difference	0.18	5.52	0.17	5.06
Random effects	***SD***		***SD***	
Word (intercept)	0.04		0.05	
Participant (intercept)	0.31		0.32	
Starting Difference by Participant (slope)	0.13		0.13	
Residual	0.29		0.26	

By removing data points with standardized residuals that differed from 0 by more than 2.5 standard deviations, the influence of outliers was further limited. As Figure 4 illustrates, refitting the model on the trimmed data (see Table 2) also resulted in more normally distributed residuals compared to the original model.

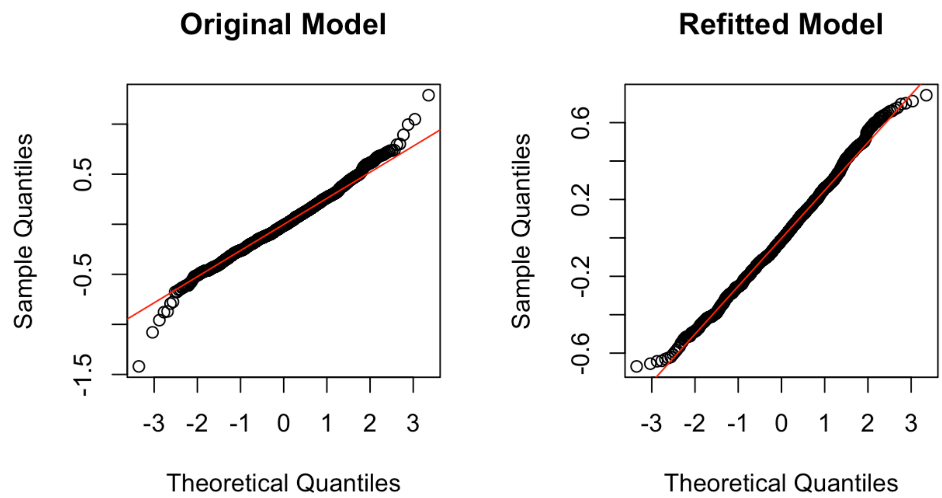

Figure 4: Quantile-quantile plots of the doubled pitch difference model's residuals before (left) and after (right) removing outliers and refitting the model.

Finally, the mean estimate of the doubled pitch difference across all data points of the trimmed dataset was calculated to be 2.22 ERB (SD = 0.46). In other words, the participants estimated that the 1.20 ERB difference between versions A and B of the words used in this experiment was perceptually doubled by a 2.22 ERB difference. The strong negative relationship between pitch adjustment and starting pitch difference, $r = -0.79$ (see also Figure 3), shows that this estimate reflects participants' auditory perceptions rather than random variation or visual cues.

3.3 Discussion

By asking participants to adjust a pitch difference to be double the size of a 1.2 ERB difference, this experiment verified whether our method of manipulation could be used to create a perceptually reliable increase in the pitch of our stimuli. The data showed that most participants were able to do so, resulting in a mean estimate of 2.22 ERB representing the perceptual doubling.

On a side note, the results are also interesting in light of the psychoacoustic scale described by formula (5). According to this scale, a difference twice the size of 1.2 ERB would correspond to 2.4 ERB. If it is assumed that this scale does indeed represent the true relation between F0 and perceived pitch, 2.4 ERB can be taken as the population mean for the sample of estimates gathered in this experiment. A one-sample t-test can then be used to show that the sample mean of 2.22 ERB is significantly different from the population mean, $t(1239) = -13.98$, $p < .001$. This difference suggests that something about the materials, experimental design, or participants is uncharacteristic of the perceptual behaviour described by the ERB scale.

This experiment has resulted in two f0 differences that correspond to increasingly large perceptual differences along the psychological dimension of pitch. Assuming the exemplar view of similarity, these f0 differences can be used to create conditions of decreasing similarity between tokens of the same word. The following section implements these conditions in an exemplar experiment.

4. Exemplar Experiment

4.1 Methods

Procedure

Participants performed a lexical decision task in response to auditory stimuli by pressing a button with the index finger of the dominant hand if a stimulus was an existing word and pressing a button with the index finger of the other hand if the stimulus was not an existing word. Response times were measured from the onset of the stimuli. With the exception of an additional between-subjects condition, a typical exemplar priming paradigm was used in this experiment. Participants were presented with prime-target pairs that were either identical or different. Two breaks divided the experiment into three blocks, each of which consisted of a prime and a target phase. Primes and their corresponding targets were separated by a minimum of 21 and a maximum of 58 other trials.

After doing the exemplar priming experiment, all subjects performed an experiment intended to measure between-participant differences in perceptual acuity regarding pitch differences. This was implemented by presenting participants with two consecutive productions of the same word and asking them whether they were identical or different in pitch. The pitch difference between the two versions was manipulated to varying degrees across trials. The number of correct responses by a participant was taken as measure of that person's perceptual sensitivity when it comes to pitch differences (henceforth this measure will be referred to as the *pitch sensitivity score*).

Materials

In addition to the 70 words used in the pitch adjustment experiment, the exemplar experiment used an equal amount of pseudo-words that followed the same phonotactic constraints as the real words. This made for a total of 140 prime-target pairs that were used to create 280 trials. All stimuli in the *same* condition as well as the target stimuli in the *different* conditions were created by shifting the original recordings down by -0.4 ERB. The prime stimuli in the *different* conditions were resynthesized at +0.8 or +1.82 ERB, resulting in 1.2 and 2.22 ERB prime-target differences for the respective between-subjects *different* conditions.

The pitch sensitivity experiment used the 10 of the 70 words in the exemplar and pitch adjustment experiments to create 140 trials. The shift in f0 between the two tokens in each trial was either 0, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.8, 1, or 1.2 ERB.

Participants

In total, 64 people were recruited from the MPI participant pool. They were paid € 6 for their cooperation. The subset of participants that was presented with the smaller pitch difference consisted of 22 females and 10 males with a mean age of 22.13 years (SD = 2.31). The group that was exposed to the larger pitch difference was made up of 23 females and 9 males with a mean age of 21.53 years (SD = 1.78). The participants presented with the small mismatch difference had a significantly higher pitch sensitivity score (M = 81.91; SD = 9.02) than the participants that encountered the large mismatch difference (M = 87.88; SD = 11.13), $t(62)$, $p = 0.022$.

4.2 Results

4.2.1 Exemplar Priming

First, all response times were logarithmically transformed in order to limit the influence of a skewed distribution on the analysis. Based on the inspection of quantile-quantile plots for each participant, data points with response times longer than approximately 1585 ms were excluded from further analysis. The remaining dataset was used to build a mixed-effects model.

Apart from the *exemplar priming* and *mismatch difference* predictors and their interaction, a number of covariates could reasonably be expected to influence response times to the target stimuli. Word duration, RT to corresponding prime, RT to previous trial, word frequency, participants' pitch sensitivity score, prime-target lag, and trial number were all considered as control variables. However, only the control variables that significantly improved the model were retained, see Table 3.

Table 3: Linear mixed model of the participants' log-transformed RTs to the target stimuli.

	Original Model		***Refitted Model***	
Fixed effects	***B***	***t***	***B***	***t***
Intercept	1.98	35.51	1.98	40.13
Exemplar Priming: Mismatch	0.004	1.60	0.005	2.13
Mismatch Difference: Large	-0.04	-4.51	-0.04	-4.67
Word Duration	0.34	11.49	0.36	12.09
RT to Prime	0.14	9.18	0.14	10.46
RT to Previous Trial	0.13	10.54	0.12	11.54
Exemplar Priming: Mismatch × Mismatch Difference: Large	-0.008	-2.06	-0.007	-2.24
Random effects	***SD***		***SD***	
Word (intercept)	0.02		0.02	
Participant (intercept)	0.03		0.03	
Residual	0.06		0.05	

Following Baayen (2008, pp. 256-257), the resulting model was used to remove data points with standardized residuals that differed from 0 by more than 2.5 standard deviations. Subsequently, the model was refitted to the remaining data points (see Table 3). As a result, the normality of the residuals was improved, as Figure 5 illustrates.

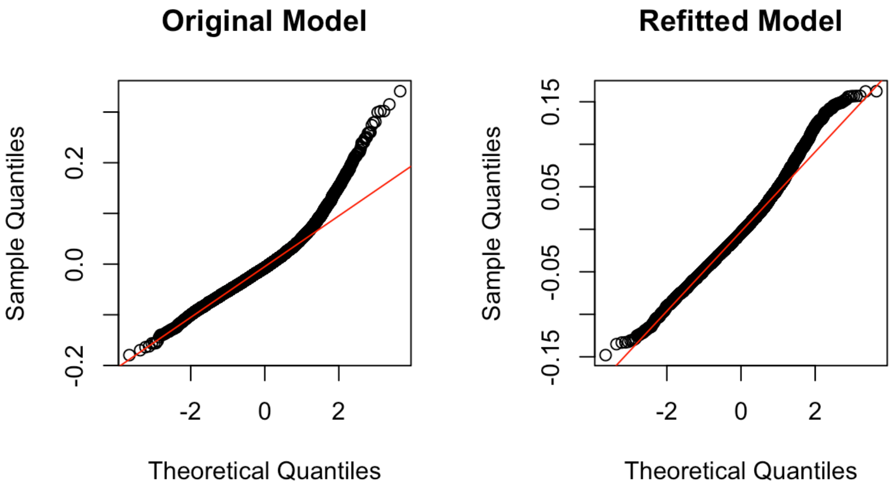

Figure 5: Quantile-quantile plots of the target RT model's residuals before (left) and after (right) removing outliers and refitting the model.

As Table 2 shows, a significant interaction between *exemplar priming* and *mismatch difference* exists. To facilitate the interpretation of this interaction, this effect has been visualized in Figure 6.

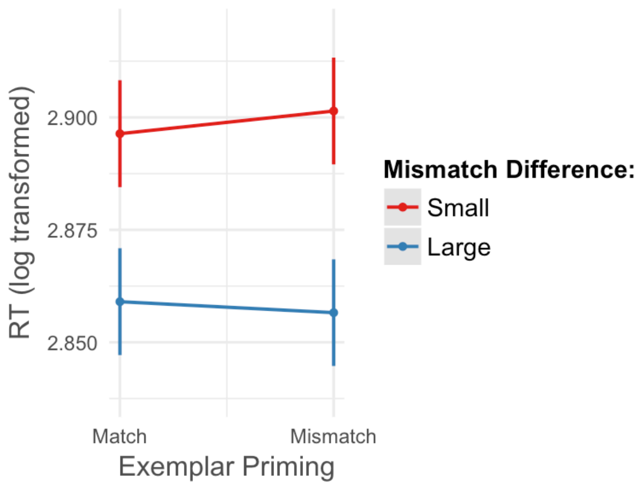

Figure 6: Predicted values for target RTs in the different conditions of the *Exemplar Priming x Mismatch Difference* interaction effect. Vertical lines represent 95% CIs.

Figure 6 reveals that the participant group presented with the smaller *mismatch difference* responded more quickly to targets that matched their respective primes in pitch compared to targets that did not. A comparison using least square means[2] showed that this difference was significant, $t(3692.11) = -2.13$, $p = .03$. In other words, these participants showed a difference in response times that could be explained as an exemplar effect. On the other hand, participants presented with the larger *mismatch difference* responded more slowly to targets that matched their respective primes in pitch compared to targets that did not. However, comparing the least

[2] All comparisons of the model's least square means used the Satterthwaite approximation to degrees of freedom.

square means showed that this difference was not significant, $t(3692.62) = 1.04$, $p = .30$. These participants, then, did not show any significant difference in response times that could be interpreted as an exemplar effect.

Furthermore, Figure 6 suggests that participants from the large *mismatch difference* group responded faster than those in the small *mismatch difference* group. Because a main effect of the *mismatch difference* variable was not hypothesized and because this variable is involved in a significant interaction, the significance of this main effect was investigated using post-hoc pairwise comparisons between all conditions in the *Exemplar Priming* × *Mismatch Difference* interaction. All of the 4 possible pairwise comparisons that involved subsets from both *mismatch difference* conditions produced significant Tukey-corrected *p*-values, all ps < .001. It follows that one between-subjects group had significantly faster responses than the other group, regardless of whether the target stimulus matched the pitch of its prime stimulus.

4.2.2 Processing of primes

Participants' response times to the prime stimuli were analysed in the same way as their RTs to the target stimuli. The resulting mixed-effects model is summarized in Table 4.

Table 4: Linear mixed model of the participants' log-transformed RTs to the prime stimuli.

	Original Model		***Refitted Model***	
Fixed effects	***B***	***t***	***B***	***t***
Intercept	2.23	40.36	2.22	44.70
Exemplar Priming: Mismatch	-0.003	-1.23	-0.004	-1.66
Mismatch Difference: Large	-0.04	-5.55	-0.04	-5.60
Word Frequency	-0.01	-4.84	-0.01	-4.61
Word Duration	0.32	9.31	0.34	10.15
RT to Target	0.12	9.10	0.12	10.22
RT to Previous Trial	0.08	6.34	0.08	6.78
Exemplar Priming: Mismatch × Mismatch Difference: Large	0.01	2.92	0.01	3.24
Random effects	***SD***		***SD***	
Word (intercept)	0.02		0.02	
Participant (intercept)	0.03		0.03	
Residual	0.06		0.05	

As with the model for target RTs, the normality of the model residuals was improved by removing data points with standardized residuals that differed from 0 by more than 2.5 standard deviations, see Figure 7.

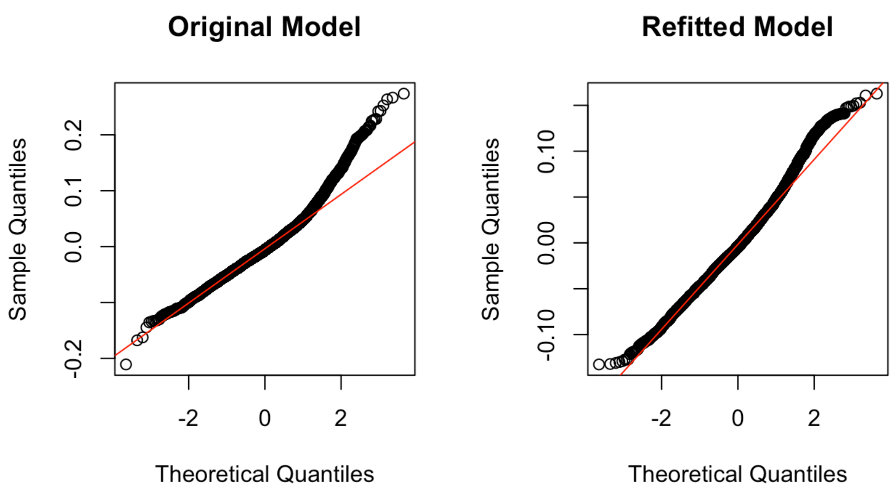

Figure 7: Quantile-quantile plots of the prime RT model's residuals before (left) and after (right) removing outliers and refitting the model.

Table 2 reveals a significant interaction between the *exemplar priming* condition and the between-subjects *mismatch difference* condition of the prime stimuli. This interaction has been visualized in Figure 8.

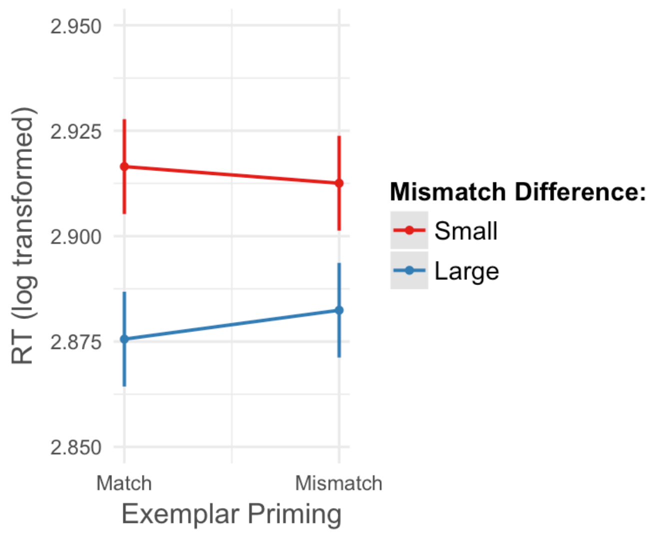

Figure 8: Predicted values for prime RTs in the different conditions of the *Exemplar Priming x Mismatch Difference* interaction effect. Vertical lines represent 95% CIs.

Post-hoc pairwise comparisons of the least square means were conducted to tease apart this interaction, as no specific hypotheses were formulated regarding the nature of a significant interaction effect on RTs to the prime stimuli. The Tukey adjusted p-values of all possible pairwise combinations revealed that, for participants in the small *mismatch difference* group, there was no significant difference between RTs to stimuli in matching versus mismatching *exemplar priming* conditions. However, for participants in the large *mismatch difference* group, RTs to stimuli in the matching *exemplar priming* condition were significantly faster than those to primes in the mismatching condition, $p = .018$.

As in the analysis of target RTs, all 4 pairwise comparisons that involved both conditions of the *mismatch difference* variable showed that responses to small *mismatch differences* were slower than those to large *mismatch differences*, all ps < .001.

4.3 Discussion

Although the analysis of RTs to the target stimuli showed a significant interaction between *exemplar priming* conditions and the size of the *mismatch difference*, the role of *mismatch difference* in this interaction was not in the hypothesized direction. The exemplar effect that was observed in the small *mismatch difference* condition was not present—let alone increased—in the large *mismatch difference* condition. It was expected that the increase in target RTs associated with a pitch difference between prime and target stimuli (relative to no prime-target pitch difference) would be even greater for prime-target pairs with a larger pitch difference. Instead, it was found that prime-target pairs with a larger pitch difference resulted in slightly (though not significantly) faster target RTs than identical prime target pairs.

Multiple explanations for these results can be considered. As the *mismatch difference* variable is between-subjects, perhaps the most straightforward explanation would be that the subjects exposed the larger prime-target pitch difference in the mismatch condition were somehow less sensitive or attentive to pitch differences. However, the results of the pitch sensitivity experiment suggest the opposite: this group performed better than the group that did show an exemplar effect.

Alternatively, the lack of an exemplar effect could be attributed to the influence of factors other than perceptual similarity on the spoken word comprehension process. In exemplar theoretic terms, the activation level of exemplars does not only depend on its similarity to the incoming stimulus. It also depends on their memory strength, which can be influenced by its salience (as discussed in Section 2.1). If this line of reasoning is applied to our experimental design, it follows that a prime stimulus that is somehow stored as a salient exemplar could facilitate processing of a target stimulus due to its high activation level, regardless of the pitch difference between prime and target. In light of our results, this raises the question whether the prime stimuli in the large *mismatch difference* condition were somehow more salient than the primes in the small *mismatch difference* condition. As mentioned in Section 4.1, the prime stimuli for the the small pitch differences were resynthesized at +0.8 ERB relative to the original, whereas the prime stimuli for the large pitch differences were resynthesized at +1.82 ERB relative to the original. It could be, then, that the higher pitch of the primes in the large *mismatch difference* condition resulted in salient exemplars. The increased memory strength of these salient exemplars could have compensated for their pitch differences with the corresponding target stimuli during the exemplar activation process. As a result, the activation level of the primes in this participant group's mismatch condition might have been comparable to that of the matching primes, which translated into similar target RTs in those conditions. Although this explanation is mostly based on theory, there is some supporting evidence from the data. The data on prime RTs showed that participants presented with +1.82 ERB shifted primes responded more slowly to those primes than to -0.4 ERB shifted primes. On the other hand, the participant group that encountered +0.8 ERB shifted primes responded equally fast to those primes as they did to -0.4 ERB shifted primes. These results could be a reflection of the salience of the +1.82 ERB shifted primes, as it is not unthinkable that stimuli with a salient pitch level are processed in a different time frame than stimuli with regular pitch levels.

Another unexpected result of this experiment concerned the difference in RT between the two conditions of the *mismatch difference* variable. In both the target and the prime data, the responses in the small *mismatch difference* condition were slower than those in the large *mismatch difference* condition. Given that the only consistent difference between these conditions are the participant groups, it is likely that differences between the two participant groups are to blame for this unexpected result. One hint at what these differences might be is provided by the pitch sensitivity results. The increased pitch sensitivity of the participants in the large *mismatch difference* group could be the result of increased attention to the experimental task. In this light, the faster responses in the priming experiment could also be interpreted as increased performance as the result of more attentive participants.

5. Conclusion

The current research has attempted to provide additional evidence for an exemplar approach to spoken word comprehension that involves the representation of phonetic detail. More specifically, it investigated to what extent the processing of a stimulus is dependent on its similarity to previously encountered tokens of that word which are stored as exemplars. A formal exemplar model was discussed in which this concept of similarity was defined as being inversely related to differences between the stimulus and the exemplars regarding their respective positions along certain psychological dimensions.

It was argued that the relationship between similarity and exemplar priming cannot be investigated in much detail, if it is not known how acoustic variation manifests itself as differences along psychological dimensions. This was illustrated in the discussion of exemplar priming experiments that manipulated similarity by comparing responses to identical prime-target pairs with responses to different prime-target pairs. Because these experiments did not control the similarity of non-identical prime-target pairs, they could not rule out the possibility that the increase in priming effect associated with exemplar matching only occurs with identical stimuli. Subsequently, research was discussed that solved this problem by applying MDS to estimate which (combination of) indexical and phonetic features could be transformed into psychological dimensions that explain priming effects between non-identical stimuli.

Although the results of the MDS study suggested that exemplar effects are not dependent on identical prime-target pairs, it was argued that the nature of that method warrants research that confirms its findings using a different methodology. As such, one of the psychological dimensions proposed by MDS, pitch, was manipulated through resynthesis to create conditions of decreasing prime-target similarity. The acoustic correlate of a pitch difference twice the size of 1.20 ERB was estimated to be 2.22 ERB in an adjustment experiment. Subsequently, these two pitch differences were used as conditions of decreasing similarity in an exemplar priming experiment. Given the hypothesis that decreased prime-target similarity results in less priming of the target, it was expected that, compared to a condition with identical primes and targets, the condition with a 2.22 ERB difference would show a larger increase in RTs than the condition with a 1.20 ERB difference. However, the results only showed a significant increase in target RTs for the smaller pitch difference. Surprisingly, the target RTs in the condition with the larger pitch difference were not significantly slower than those in those in the condition without a pitch difference between primes and targets. In other words, the comparison of conditions that was expected to show the biggest exemplar effect did not show an exemplar effect at all. In light of this result, it was suggested that the pitch manipulation of the prime stimuli used for the larger pitch difference was salient to the participants, which could have compensated for the decrease in activation of the prime exemplars during processing of the target stimuli.

In conclusion, the pitch adjustment experiment successfully produced estimates of the acoustic correlates of a pitch doubling. However, the inconsistent results of the exemplar priming experiment highlight the need to control all factors that might influence the activation level of exemplars. In this light, future research would do well to investigate the memory strength variable and the factors that contribute to it.

References

Baayen, R.H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341-345.

Church, B.A., & Schacter, D.L. (1994). Perceptual specificity of auditory priming: implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(3), 521-533.

Craik, F.I.M., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology, 26*(2), 274-284.

Drager, K. & Kirtley, J. (2016). Awareness, salience, and stereotypes in exemplar-based models of speech production and perception. In A. M. Babel (Ed.), *Awareness and Control in Sociolinguistic Research* (pp. 1-24). Cambridge: Cambridge University Press.

Ehrenstein, W.H., & Ehrenstein, A. (1999). Psychophysical methods. In U. Windhorst & H. Johansson (Eds.), *Modern Techniques in Neuroscience Research* (pp. 1211-1241). Berlin: Springer-Verlag

Ernestus, M. (2014). Acoustic reduction and the roles of abstractions and exemplars in speech processing. *Lingua, 142*, 27-41.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London: SAGE.

Forbach, G.B., Stanners, R.F., & Hochhaus, L. (1974). Repetition and practice effects in a lexical decision task. *Memory & Cognition, 2*(2), 337-339.

Forster, K.I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(4), 680-698.

Goldinger, S.D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1166-1183.

Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251-279.

Greenwood, D. D. (1990). A cochlear frequency-position function for several species—29 years later. *The Journal of the Acoustical Society of America, 87*(6), 2592-2605.

Hanique, I., Aalders, E., & Ernestus, M. (2013). How robust are exemplar effects in word comprehension? *The Mental Lexicon, 8*(3), 269-294.

Hermes, D.J., & Gestel, J.C. van. (1991). The frequency scale of speech intonation. *The Journal of the Acoustical Society of America, 90*(1), 97-102.

Hintzman, D.L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*(4), 411-428.

Johnson, K. (2005). Decisions and mechanisms in exemplar-based phonology. In *UC Berkeley Phonology Lab Annual Report 2005* (pp. 289–311).

Kruskal, J.B., & Wish, M. (1978). *Multidimensional scaling*. Newbury Park, CA: SAGE.

Longster, J.A. (2003). *Concatenative speech synthesis: a Framework for Reducing Perceived Distortion when using the TD-PSOLA Algorithm.* Unpublished doctoral dissertation, Bournemouth University,

Luce, P.A., & McLennan, C.T. (2005). Spoken word recognition: The challenge of variation. In D.B. Pisoni & R.E. Remez (Eds.), *Handbook of Speech Perception* (pp. 591-609). Malden, MA: Blackwell.

Medin, D.L., & Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review, 85*(3), 207-238.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech communication, 9*(5-6), 453-467.
Nijveld, A., Bentum, M., Bosch, L. ten, & Ernestus, M. (2016). *The occurrence of exemplar effects is better predicted by token match than by experimental task*. Manuscript in preparation.
Nijveld, A., Ten Bosch, L., & Ernestus, M. (2015). Exemplar effects arise in a lexical decision task, but only under adverse listening conditions. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow: University of Glasgow.
Nolan, F. (2003, August). Intonational equivalence: an experimental evaluation of pitch scales. In *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona* (Vol. 39).
Nosofsky, R.M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 104-114.
Nosofsky, R.M. (1991a). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17*(1), 3-27.
Nosofsky, R.M. (1991b). Typicality in logically defined categories: Exemplar-similarity versus rule instantiation. *Memory & Cognition, 19*(2), 131-150.
Nosofsky, R.M., & Johansen, M.K. (2000). Exemplar-based accounts of “multiple-system” phenomena in perceptual categorization. *Psychonomic Bulletin & Review, 7*(3), 375-402.
Pierrehumbert, J.B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J.L. Bybee & P.J. Hopper (Eds.), *Frequency and the Emergence of Linguistic Structure* (pp. 137-158). Amsterdam: John Benjamins.

Appendix A: Categorization Bias in a Within-Subject Design

Some exemplar models allow for gradient encoding of physical properties along the hypothesized psychological dimensions (e.g. Nosofsky, 1984), whereas other models use categorical encoding (e.g. Hintzman, 1986). The level of granularity at which people encode physical characteristics of stimuli is not known. If participants encode the fundamental frequency of lexical stimuli categorically, a bias might arise when more than two categories are involved. In order to illustrate this bias, a hypothetical experiment that involves three categories will now be considered.

Suppose the pitch levels of the stimuli in this experiment can be described by three underlying distributions. The first distribution is centred around a pitch value of 1, the second distribution has a mean pitch of 2, and the third distribution has an average pitch of 3. In this example, it is assumed that participants categorize the pitch value of every stimulus from these distributions into pitch category A, B, or C. It is assumed that categorization is probabilistic and is based on the similarity between the pitch value of the stimulus and the mean pitch values of the categories. For convenience sake, it is assumed that similarity s is calculated from the absolute difference between the mean pitch of category A and the stimulus i_1, see 6:

$$s_{Ai} = |P_A - P_i|^{-1} \tag{6}$$

The similarity of a stimulus i to the respective categories can then be used to calculate the respective probabilities that it is sorted in each category. For instance the probability that stimulus i is sorted into category A is calculated in 7:

$$P(A|i) = \frac{s_{Ai}}{s_{Ai} + s_{Bi} + s_{Ci}} \tag{7}$$

Applying this categorization mechanism to the three incoming stimuli in Figure 9 results in the probabilities of being sorted into the respective categories in Table 5.

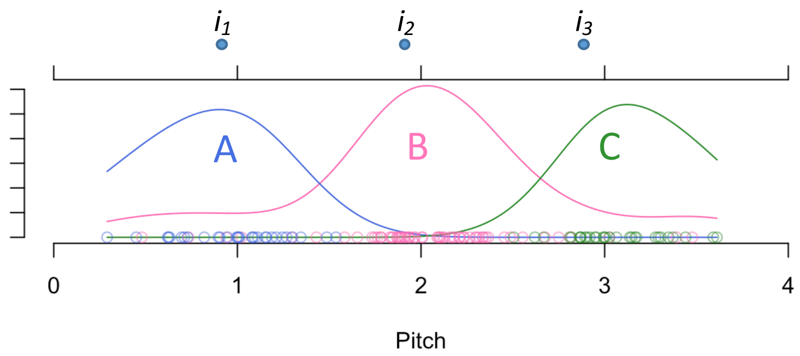

Figure 9: Visualization of incoming stimuli and pitch categories.

Table 5: Calculation of categorization probabilities for stimuli in Figure 9.

		Incoming Stimuli			Category Probabilities	
		$i_1 = 0.9$	**$i_2 = 1.9$**	**$i_3 = 2.9$**	**Raw**	**Adj.**
Categories	**A=1**	$\frac{10}{10 + 0.91 + 0.48}$	$\frac{1.11}{1.11 + 10 + 0.91}$	$\frac{0.53}{0.53 + 1.11 + 10}$	0.339	0.332
	B=2	$\frac{0.91}{10 + 0.91 + 0.48}$	$\frac{10}{1.11 + 10 + 0.91}$	$\frac{1.11}{0.53 + 1.11 + 10}$	0.336	0.336
	C=3	$\frac{0.48}{10 + 0.91 + 0.48}$	$\frac{0.91}{1.11 + 10 + 0.91}$	$\frac{10}{0.53 + 1.11 + 10}$	0.326	0.332

Table 5 shows the mean probabilities that an incoming item is sorted into the respective categories. The *Raw* column shows these probabilities for the stimuli in Figure 9, which are shifted down by -0.1 pitch relative to the category means. If we correct for this -0.1 shift by doing the same probability calculations for stimuli that are shifted up +0.1 pitch and taking the average of both mean probabilities, we get the *Adjusted* category probabilities in the rightmost column. These probabilities show that there is a bias towards category B. This bias towards the middle category would be amplified if probabilities are weighted by the number of items in each category.