Available online at www.sciencedirect.com
ScienceDirect
Journal of Phonetics 34 (2006) 458–484
Journal of Phonetics
www.elsevier.com/locate/phonetics

Factors influencing speech perception in the context of a merger-in-progress

Jennifer Hay[a,*], Paul Warren[b], Katie Drager[a]

[a]University of Canterbury, Christchurch, New Zealand
[b]Victoria University of Wellington, Wellington, New Zealand

Received 5 January 2005; received in revised form 29 June 2005; accepted 4 October 2005

Abstract

In New Zealand English there is a merger-in-progress of the NEAR and SQUARE diphthongs. This paper investigates the consequences of this merger for speech perception.

We report on an experiment involving the speech of four New Zealanders—two male, and two female. All four speakers make a distinction between NEAR and SQUARE. Participants took part in a binary forced-choice identification task which included 20 NEAR/SQUARE items produced by each of the four speakers. All participants were presented with identical auditory stimuli. However the visual presentation differed. Across four conditions, we paired each voice with a series of photos—an "older" looking photo, a "younger" looking photo, a "middle class" photo and a "working class" photo. The middle and working class photos were, in fact, photos of the same people, in different attire. In a fifth condition, participants completed the task with no associated photos. At the end of the identification task, each participant was recorded reading a NEAR/SQUARE wordlist, containing the same items as appeared in the perception task.

The results show that a wide range of factors influence accuracy in the perception task. These include participant-specific characteristics, word-specific characteristics, context-specific characteristics, and perceived speaker characteristics. We argue that, taken together, the results provide strong support for exemplar-based models of speech perception, in which exemplars are socially indexed.

1. Introduction

Merger is a process, much studied by sociolinguists (e.g. Labov, 1994; Gordon, 2002), in which sound change leads to the collapse of a phonemic contrast, so that what were previously two distinct phonemes in a dialect come to be realized as a single phoneme.

In New Zealand English there is a merger-in-progress of the NEAR and SQUARE[1] vowels which has been proceeding through the variety over the last few decades (Bayard, 1987; Holmes, Bell, & Boyce, 1991; Maclagan & Gordon, 1996; Gordon & Maclagan, 2001; Batterham, 2000). New Zealand English is

*Corresponding author. Tel.: +64 3 364 2242; fax: +64 3 364 2969.
E-mail address: jen.hay@canterbury.ac.nz (J. Hay).

[1]We use Wells' (1982) lexical set words to refer to the vowels under discussion.

0095-4470/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wocn.2005.10.001

non-rhotic,[2] and for all speakers these vowels are diphthongs with schwa offglides. For distinct speakers, NEAR (/iə/) has a closer first element than SQUARE (/eə/). The initial direction of the merger was unclear, with some speakers merging on the closer variant, and some on the more open variant. However for most younger speakers the merger is now near complete on the NEAR variant (with a typical realization of [i̟ə]).

In Warren, Hay, and Thomas (forthcoming) we report on a series of experiments designed to investigate the consequences of this merger for speech perception. Taken together, the results provided evidence in favor of experience-based models of spoken word recognition, with both lexical and pre-lexical levels. In such models, the representation of a particular word is a distribution of remembered exemplars, complete with phonetic detail. Warren et al. (forthcoming) provided preliminary evidence that the set of prior experiences must also include implicit knowledge of how variation is socially distributed. This paper describes an experiment designed to test this explicitly.

2. Background

A comprehensive survey of the NEAR-SQUARE merger and its progress over the last twenty years or so is provided by Gordon and Maclagan (2001), who present data from a long-term study of 14–15 year old students in Christchurch. These authors have re-visited the same schools every 5 years since 1983, obtaining recordings of words containing NEAR and SQUARE vowels, read in sentence contexts and word-lists.[3] The distinction between the diphthongs was still widely present in the first recordings in 1983, and these earliest samples also show considerable variation, with some speakers showing a lack of distinction, but no clear pattern of merger towards either NEAR or SQUARE. By 1998, however, there was an almost complete merger on NEAR. Gordon and Maclagan (2001, p. 232) thus describe the merger as a "merger of approximation" rather than a "merger of expansion" (Labov, 1994, p. 321). That is, the diphthongs are collapsing on a single form, rather than using the whole range of pronunciations previously attested for both NEAR and SQUARE. Maclagan and Gordon (1996) supplement their long-term study with an apparent-time comparison of two age groups who they recorded in 1994—the younger speakers were 20–30 years old and the older speakers were 45–60 years. The younger speakers again show a more complete merger towards NEAR than the older speakers.

It has also been claimed that the Christchurch survey shows the NEAR-SQUARE merger moving through NZE by a process of lexical diffusion, affecting some words before others (Maclagan & Gordon, 1996, pp. 131–133). Additional data collected in the early 1990s in Porirua, just north of Wellington, by Holmes and Bell (1992) were subjected by them to auditory analysis. Warren (2005) has re-analyzed these data and looked in particular at the preceding phonetic context for the materials used (i.e. the consonant before the NEAR or SQUARE vowels). It transpires that the nature of the preceding consonant may also contribute to the pattern of diffusion. The original study showed that SQUARE-raising increased over apparent time, with mid-age speakers showing closer first elements than older speakers, and young speakers producing the closest first elements of all three groups. The reanalysis of the data showed that this raising was present for all preceding phonetic contexts for the youngest speakers, but only after coronal consonants for the mid-age speakers. There is a suggestion then in these data that at an early stage the change may have been conditioned by the place of articulation of the preceding consonant—the first element of the diphthong would be a natural consequence of coarticulation with a preceding coronal consonant. The subsequent raising of SQUARE in other phonetic contexts provides a pattern that fits with Ohala's (1992) suggestion that some sound changes are a consequence of listeners' failure to compensate for coarticulation. That is, NZE speakers may have "forgotten" that a conditioning factor determined the closer onset for SQUARE vowels after coronal consonants. These post-coronal SQUARE vowels were consequently reinterpreted as NEAR vowels, and this reinterpretation then spread to other words formerly spoken with SQUARE vowels.

[2]With the exception of speakers from small region at the bottom of the South Island, who produce rhotic forms of words in the NURSE lexical set (e.g. *work*, *third*, *hurt*, etc.).

[3]While read (rather than spontaneous) materials are not ideal, Gimson (1963, p. 143), referring to the original study by Fry (1947), lists the SQUARE and NEAR vowels as only the 17th and 18th most frequent out of 20 English vowels respectively. It, therefore, becomes necessary to use read materials in order to elicit sufficient tokens for analysis.

As a sound change such as the merger of NZE NEAR and SQUARE advances, patterns of variation in pronunciation will change. If the initial impetus for change was—as indicated above—within the SQUARE set, then to start with we would expect to see more variation within this set, with a range of close as well as more open first elements for the diphthong. The asymmetric merger of approximation on NEAR means that as the change spread an increasing number of SQUARE targets received NEAR pronunciations. While the change is still incomplete, and continues to advance through the speech community, some speakers (in this case the older and/or more conservative speakers) will still maintain a NEAR-SQUARE distinction, while others (mainly younger speakers) will on the whole produce only the NEAR form. Variation within the community will therefore be speaker dependent, but overall there will be greater variation in the realization of SQUARE than in that of NEAR. An interesting question is whether listeners are able to utilize their knowledge of speaker differences in order to help interpret the variation that they hear (Johnson, Strand, & D'Imperio, 1999; Strand, 1999; Drager, 2005b). Psycholinguistic investigation of the consequences of the merger (Warren & Hay, 2005; Warren, Rae, & Hay, 2003) shows that young New Zealanders are likely to respond to NEAR forms as tokens of either NEAR or SQUARE words (i.e. they will treat NEAR forms as lexically ambiguous between such words, so that [tʃiə] could be understood as either a vocal exclamation or an item of furniture) but will respond to SQUARE forms only as tokens of SQUARE words (so [tʃeə] is only the item of furniture). Access to different lexical meanings was shown in a lexical decision task with semantic priming, in which response times to words like *sit* and *shout* were measured when these words were heard as next items in stimulus lists after either [tʃiə] or [tʃeə].[4] The result suggests that young New Zealanders have not totally lost their sensitivity to SQUARE forms, which makes sense since they will after all hear these from their parents and grandparents.

In Warren et al. (forthcoming) we added a series of identification experiments to the psycholinguistic data, and we also collected production data from the same participants. In these studies, we found that accuracy in the perception task is positively correlated with the extent to which participants keep NEAR and SQUARE tokens distinct in their own speech. We also found that both production and perception scores are strongly correlated with a measure of social class based on the Elley–Irving scale (Elley & Irving, 1985). Participants from higher socio-economic backgrounds were more likely to produce distinct forms of the diphthongs. In Warren et al. (forthcoming) we argue that participants are most likely to associate with speakers from a similar background to themselves, and so those from a higher socio-economic group would have had greater exposure to distinct NEAR and SQUARE forms, giving them a greater sensitivity to the difference in the identification task. These findings regarding social class tie in with Maclagan and Gordon's (1996) school study, where they report that "the merger seems to be more complete for the lower socioeconomic classes than for the higher ones" (Maclagan & Gordon, 1996, p. 136).

In our previous study we also found considerable variation in the error rates for different items. Interestingly, we found that item error rates correlated not with the extent of the acoustic difference between the members of a stimulus pair as heard by participants in the identification task, but with the extent to which the pair members were kept distinct in the production data from these participants. This result seems again to suggest that listeners' sensitivity to NEAR-SQUARE differences depends on the extent to which they have experienced these as distinct. This experience depends on the distinctiveness of individual items as well as on the company that the participants keep.

In further experiments, Warren et al. (forthcoming) measured error rates in identification tasks with old and young male and female voices. These error rates gave an indication of how speaker variables might affect identification. There was an overall effect of the perceived age of the voice, such that most errors were in response to younger voices. This is compatible with the general tendency in previously reported production data (Holmes & Bell, 1992; Maclagan & Gordon, 1996) for young speakers to be more advanced in the merger. There was also a consistent finding (which we have found also in our other identification tasks) of more errors in response to SQUARE tokens than for NEAR tokens.

Our overall analysis of the results in Warren et al. (forthcoming) was that they provided evidence for a model of perception and word recognition that includes the following features: a fast pre-lexical processor

[4]While [ ] is used here to indicate phonetic tokens, the transcription should not be interpreted as providing an absolutely precise indication of the quality of the vowel concerned.

(Pierrehumbert, 2001a), lexical storage of phonetic exemplars (Hawkins, 2003; Johnson, 1997; Pierrehumbert, 2002; Pitt & Johnson, 2003), and social indexing of these exemplars (Johnson et al., 1999; Strand, 1999).

The fast pre-lexical processor—in agreement with other findings concerning the relative robustness of low and high frequency items in speech perception (Hay, Pierrehumbert, & Beckman, 2003; Newman, Sawusch, & Luce, 2000; Savin, 1963)—is biased towards the recognition of phonemes that it has encountered most frequently. The fact that the NEAR tokens used in our experiments tend to have higher lexical frequency than the SQUARE tokens, together with the fact that the merger is towards NEAR, means that /iə/ is by far the more frequently encountered of the two vowels. So the interpretation of a heard [eə] as a NEAR vowel, due to the stronger center of gravity for /iə/, is predictable.

The storage of phonetic exemplars at the lexical level is supported by our findings that those participants who have had most exposure to distinct NEAR and SQUARE forms, because of the socio-economic group with which they associate, were best able to discriminate these forms in the identification task; that the [iə] forms gave access to both NEAR and SQUARE lexical representations in the priming task, while [eə] forms did not activate NEAR representations; and that there were higher error rates in the identification task for items which are most merged.

The social indexing of these exemplars is supported by the finding that the overall error rates and the error rates for both SQUARE and NEAR were lowest for the older speakers, and by the finding that NEAR vowels were more likely to be mis-identified when the speaker is female. It is also supported by the finding from a further semantic priming study using an older speaker, which does not find the asymmetry in priming reported by Warren and Hay (2005) and by Warren et al. (2003). In other words, the differential priming effects of NEAR and SQUARE tokens depend on perceived characteristics of the speaker.

In summary, Warren et al. (forthcoming) provided preliminary evidence that the set of prior experiences upon which speech perception and word recognition is based must also include implicit knowledge of how variation is socially distributed. In the current paper we describe an experiment designed to test this explicitly.

Strand (1999) and Johnson et al. (1999) have demonstrated that listeners' impressions of talker gender affect the location of phoneme boundaries—both for vowels and for consonants. Drager (2005a) provides evidence that the perceived age of a voice can affect the location of vowel boundaries between vowels undergoing change in New Zealand English. Drager's work also indicates that perceived social class will influence the perceived location of vowel boundaries.

In this paper, we aimed to add to this existing evidence for the effect of perceived speaker identity on speech perception, by manipulating listeners' expectations about voices producing NEAR and SQUARE items. When asked to distinguish between NEAR and SQUARE items, will participants' accuracy be affected by the perceived age or social class of a speaker? In order to explore this possibility, we conducted a NEAR/SQUARE identification task, using natural speech, and attempted to influence participants' perceptions of the stimulus voices by pairing each voice with different photos across different conditions.

3. Methodology

3.1. Overall design

Participants took part in a two-alternative forced choice perception task, in which they listened to a recording of a word, and then chose which word they thought they heard from two alternatives on a computer screen. The experiment was constructed using MediaLab on a PC laptop. The auditory stimuli contained four voices, all of which produced a reliable distinction between NEAR and SQUARE. The voices were not blocked, and the order of the tokens was pseudo-randomized. All participants were presented with identical auditory stimuli. However the visual presentation differed. Across four conditions, we paired each voice with a series of photos—an "older" looking photo, a "younger" looking photo, a "middle class" photo and a "working class" photo. The middle and working class photos were, in fact, photos of the same people, in different attire. In a fifth condition, participants completed the task with no associated photos. Following the perception task, participants were recorded reading minimal pair word lists.

3.2. Auditory stimuli

In order to construct the auditory stimuli we recorded two male and two female New Zealand English speakers who produce a reliable distinction between NEAR and SQUARE. Each speaker produced multiple tokens of a set of 10 NEAR/SQUARE minimal pairs. From these recordings we chose a single stimulus token for each word from each speaker, attempting to match, as far as possible, the length, amplitude, loudness, and the degree of distinction across different word pairs. There were a total of 80 stimulus items (four voices producing 20 words each). The word pairs included in the experiment are shown in Table 1.

Because NEAR and SQUARE are both relatively infrequent, the set of potential minimal pairs to select from was relatively limited. We avoided <ear> spellings which belong to the SQUARE lexical set (e.g. <*bear*>, <*pear*>), as earlier work has demonstrated that these are much more often interpreted as belonging to the NEAR class than forms with spellings which are unambiguously SQUARE (e.g. <*bare*>, <*pair*>) (Hay & Maclagan, 2002). It was impossible to control for lexical frequency, and so instead, we explicitly investigated the role of lexical frequency in our statistical analysis.

3.3. Visual stimuli

The participants were divided into five experimental groups. The visual stimuli for each group varied.

3.3.1. Group 1

In group 1 participants were exposed only to auditory stimuli and not to any visual stimuli.

3.3.2. Groups 2 and 3

The visual stimuli for groups 2 and 3 consisted of four different photographs, two male and two female. The ages of the people in the photographs varied so that there was a photograph of an older female (OF), an older male (OM), a younger female (YF), and a younger male (YM). Examples of the photos are shown in Fig. 1 (for OM and YF).

The photos were all head and shoulder shots, and the style of clothing was intended to be similar across all photographs. A plain background was used. The photos were associated with different gender-appropriate voices across the two groups (see Table 5).

3.3.3. Groups 4 and 5

Eight photographs were used as the visual stimuli for groups 4 and 5. There were two photographs of each of four individuals, and in each photograph the individual was dressed differently. The intention was to create one photograph of each person that looked like a middle-class speaker while in a different photograph that same individual would look like a working class speaker. In all cases we attempted to manipulate this with dress, and in some cases, the background location of the photo was also manipulated. Two of the people who were photographed were male and two were female, and the individuals in the photographs were of a similar

Table 1
Minimal pairs used in perception task

air	ear
bare	beer
dare	dear
fare	fear
hair	hear
mare	mere
pair	peer
rarely	really
share	sheer
spare	spear

Fig. 1. Examples of photographs presented to groups 2 and 3: Older male and younger female.

Fig. 2. Photographs of one of the females presented to groups 4 and 5, with varying social class. The photo on the left shows the working class guise, and the photo on the right shows the middle class guise.

Fig. 3. Photographs of one of the males presented to groups 4 and 5, with varying social class. The photo on the left shows the working class guise, and the photo on the right shows the middle class guise.

age to each other. Example photographs are shown in Figs. 2 and 3. The two different photos of the same individual were associated with the same voice in the two groups (see Table 5).

3.4. The perceived social characteristics of the voices and photos

We wanted to verify the degree to which perceived age and social class were successfully manipulated with the photos. We also wanted to establish whether the four stimulus voices varied significantly in perceived age and/or social class.

For the photos, a subset of participants completed a questionnaire that would determine the extent to which the social class manipulation was successful, as well as to assess the perceived age of the individuals in the

photos. 22 participants provided ratings for the photos used in condition 4, and 24 for the photos used in condition 5. The photos used in conditions 2 and 3 were rated by 17 participants.

The rating questionnaire used questions taken from the Evaluating English Accents Worldwide project (see, e.g. Bayard, Weatherall, Gallois, & Pittam, 2001), as follows:

1. Do you recognize this person? YES NO

2. What age group would you estimate the person belongs to?
(a) 18–25 (b) 26–35 (c) 36–45 (d) 46–55 (e) 56–65 (f) 65+

3. This person gives the impression of being:

	Not at all					Very
Reliable	1	2	3	4	5	6
Ambitious	1	2	3	4	5	6
Humorous	1	2	3	4	5	6
Friendly	1	2	3	4	5	6

4. I would guess the person's educational level is:
(a) primary school (b) secondary school (c) polytechnic
(d) undergraduate degree (e) graduate/professional degree

5. The occupational area I would most expect to find this person in is:
(a) **unskilled laborer**, e.g. fast-food employee
(b) **skilled laborer**, e.g. machine operator
(c) **clerical worker**, e.g. bookkeeper
(d) **manager**, e.g. business executive
(e) **professional**, e.g. lawyer

For each photo, an average age was calculated by averaging over the midpoint of the respondents' answers to question 2 (e.g. if someone responded (b), they were taken to have answered "30.5"). Perceived educational level was calculated by transforming (a)–(e) to (1)–(5), and calculating an average for each photo. The same was done for occupation. For the purposes of statistical analyses, we have added the occupation and the education score for each individual together, to obtain an overall measure of 'social class'.

Table 2 shows the perceived age, occupation and education for the four individuals in groups 4 and 5, in their two guises. In all cases, the 'lower' class photo was rated as lower both in perceived occupation and education than the 'upper' class photo. Interestingly, individuals were consistently rated as older in their upper class guise.

Table 3 shows the perceived characteristics of the 'older' and 'younger' photos used for groups 2 and 3. Here, too, the older individuals were rated higher on the social class indices than the younger individuals.

Table 2
Estimated age, occupation and education level for the four individuals presented to groups 4 and 5 in their 'upper' and 'lower' guises

	M1		F1		M2		F2	
	Upper	Lower	Upper	Lower	Upper	Lower	Upper	Lower
Age	23.6	21	36.4	31	27.4	26.4	23.9	23.6
Occupation	4	3.6	4.4	3.1	4.1	3	4.1	3.2
Education	4	3	3.9	2.4	3.4	2	3.8	2.4

Table 3
Estimated age, occupation and education level for the four photos presented to groups 2 and 3

	Older		Younger	
	Male	Female	Male	Female
Age	48.2	54.7	22.9	23.5
Occupation	4.5	4	3.3	3.9
Education	3.9	3.7	2	3.1

Table 4
Estimated age, occupation and education level for the four stimulus voices

	Age	Education score	Occupation
M-voice-1	30.4	4.1	3.7
M-voice-2	37.2	4.3	4
F-voice-1	33.6	4	3.5
F-voice-2	41.9	3.7	3.2

Table 5
Combinations of voices and photos across different participant groups

	Group 1	Group 2	Group 3	Group 4	Group 5
M-voice-1	No photo	Old-M	Young-M	Upper-M1	Lower-M1
M-voice-2	No photo	Young-M	Old-M	Lower-M2	Upper-M2
F-voice-1	No photo	Young-F	Old-F	Lower-F1	Upper-F1
F-voice-2	No photo	Old-f	Young-F	Upper-F2	Lower-F2

Finally, similar information was collected for the four voices. These voices were played to 15 individuals who were not involved in any other part of the experiment. The ratings are based on short introductory utterances taking the form "hello, my name is _____". No NEAR/SQUARE tokens were included in the utterances. Each speaker was included on the stimulus tape twice (introducing themselves with different names). The responses to the two utterances have been averaged. The perceived age and social characteristics of the voices are given in Table 4. These values were included in the statistical analysis of the perception task discussed below, but proved not to be significant predictors of participants' behavior.

Table 5 summarizes the distributions of the photos, indicating which photo was paired with which voice for the different groups of participants in the perception experiment.

3.5. Instructions

The instructions given to participants were as follows:

> This is an experiment about similar sounding words. You will be played a series of words, and for each word, you will be asked to indicate which word you think was said. The words in this experiment are spoken by four different speakers.
>
> Because it can sometimes be difficult to listen to several different voices in one experiment, we will help you by showing you photos of the speakers. Before you hear each word, you will see a photo. This will indicate which voice to expect next. Look at the photo and then listen carefully to the recording of the speaker

producing a word. Once the word has been played, you will be given a choice of two possible words. Your task is to indicate what word you think the speaker said.

This task may not be easy. Don't worry if you're not sure about your answer, just provide your best guess. Don't think too hard about your answer—there are no right or wrong answers, and we are just interested in your first intuition. Once you have provided an answer for a word, the photo of the next speaker will be displayed, and the next word will be played.

Due to experimenter error, these instructions were given even to participants in group 1, who in fact saw no photos. Interestingly, none of the participants involved ever commented on the lack of photos in the experiment.

4. Participants

Participants were recruited from the University of Canterbury, and all of those who were analyzed were native speakers of New Zealand English. Some participants were volunteers from introductory linguistics classes, and others were recruited by advertisement, and paid $5 for their time. All participants received chocolate fish in thanks for their participation.

71 participants completed the experiment—17 in group 1, 13 in each of groups 2 and 3, and 14 in each of groups 4 and 5. In order to provide a crude measure of social class, the occupations of their parents were recorded. The corresponding scores on the New Zealand Socio-economic Index were calculated, and the higher of the parents' scores was assigned to the participant. The New Zealand Socio-economic Index (NZSEI) assigns socio-economic scores to occupations on the basis of the education and income profiles of New Zealanders. This index ranges from 0 to 100 (Davis, McLeod, Ransom et al., 1997; Davis, Jenkin, & Coope, 2003).

Table 6 shows the distribution of the 71 participants across the different groups, including the numbers of males/females, the median age in each group, and the median assigned score for the NZSEI. It can be seen that there are more females than males in the sample, and that participants in group 3 are slightly younger than in other groups.

4.1. Participants' production

The degree of the merger varies greatly across individuals. In order to determine the degree to which an individual's production was linked to their perception, production data was collected for all participants. The production task was recorded onto a tape using a Sony TCM-5000EV portable cassette-corder and a Sony ECM-F8 electret condenser microphone. The recordings were later digitized at 44.1 kHz in order to perform acoustic analysis.

Following the perception task, participants were recorded reading three minimal pair lists, in which the ten minimal pairs (Table 1) appeared in varying orders. There were no delays or distractor tasks between the three readings. In List 1, the words were given in pairs, and the order of items within pairs varied, so that sometimes

Table 6
Distribution of participants across the groups

	Group 1	Group 2	Group 3	Group 4	Group 5
Total #	17	13	13	14	14
#F	10	9	13	9	8
#M	7	4	0	5	6
Median age	24	23	19	26.5	23
(age range)	(18–50)	(18–50)	(18–32)	(18–54)	(18–47)
Median NZSEI	63	63	65	58.5	62.5
(range)	(45–90)	(25–77)	(49–77)	(30–90)	(38–90)

the NEAR token would be positioned on the left and sometimes the SQUARE token would be found on the left. The minimal pairs in List 2 were the same words as in List 1, however in List 2 the SQUARE token was always located on the right and the NEAR token was always situated on the left.

For their third reading of the word pairs, participants were asked to read the same word-list as in List 2, and were asked to comment on whether or not they thought the two words in each pair were pronounced the same in their own speech.

Because one of the primary factors conditioning people's performance in the perception task was likely to be the degree of merger in their own speech, we wanted to extract a crude measure of this from the production recordings. Similarly, we also wanted an indication of the degree of merger of different word pairs, when assessed across the entire participant pool.

In order to achieve this, all recordings were analyzed using Praat acoustic analysis software. F1 and F2 measurements were taken at the point of highest F2 during the first element of the diphthong (cf. Watson, Harrington, & Evans, 1998, p. 191). Formant values for F1 and F2 were then transformed to a Bark Scale.

In order to assign to each participant a single number which reflected the degree to which they kept the distributions distinct we calculated the Pillai score, using the R statistical analysis environment. Higher Pillai scores indicate greater distance (in F1 and/or F2) between the two vowels. Alternatively, the lower the Pillai score, the more advanced the merger. As a summary of the degree to which two distributions are kept distinct, this is superior to taking Euclidean distances between means. This is because it takes account of the degree of overlap of the entire distribution (Olson, 1976).

As lists 2 and 3 provided the words in a fixed order, it would be possible to use a strategy in order to keep the words distinct. We therefore restricted this calculation to List 1, which we guessed provided the most accurate indication of the degree to which a speaker keeps the distributions distinct in natural speech. We did also test the degree to which the overall Pillai scores predicted individuals' performance in the perception task, but found the Pillai score for List 1 to be the better predictor.

Fig. 4 shows the vowel plots for List 1 for the speaker with the highest Pillai score (0.969—a 23 year old male with a social class score of 90), and the speaker with the lowest Pillai score (0.0009—28 years, female, social class = 73). We also calculated Pillai scores for each word pair, in an attempt to quantify the degree to which the word pairs were kept distinct by the speakers we recorded. When calculated over all participants, all word pairs were kept significantly distinct ($p<0.01$). The Pillai scores ranged from 0.06 (for *really/rarely*), to 0.24 (for *beer/bare*).

Clearly a distance metric based on F1 and F2 measurements at a single point in the vowel does not capture all the potential ways in which the phonemes could potentially be distinct. It is possible there may be differences in the diphthong trajectories, durations, or voice quality (see, e.g. Di Paolo & Faber, 1990; Gordon, 2002), and detailed analysis of the production data is the subject of ongoing work. However, for the purpose of analyzing the perception experiment, we regard this metric to be an adequate indication of the degree of merger in the first element—which has been the focus of previous work on the production of these vowels.

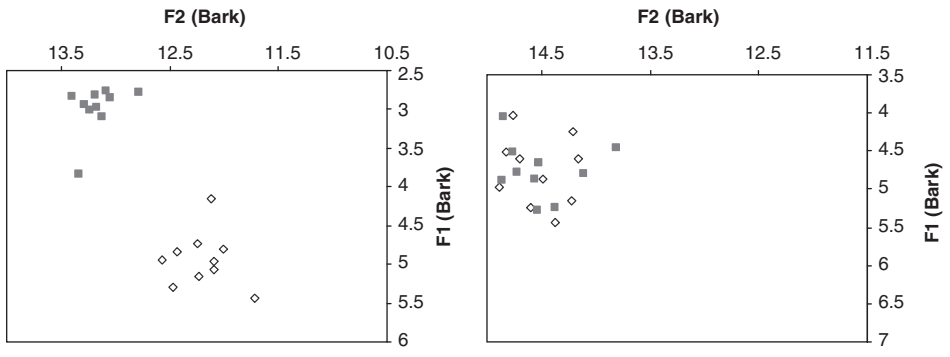

Fig. 4. Vowel plots for the most distinct speaker (male, pillai = 0.969) and least distinct speaker (female, pillai = 0.0009). Filled points represent NEAR tokens, Unfilled points represent SQUARE tokens.

4.2. Experimenters

There were two experimenters who met with the participants. One of these was a speaker of New Zealand English, and the other is from California, and speaks a variety of US English. Both experimenters maintain a distinction between the vowels in their own speech, although this distinction is much greater in the US experimenter's speech. Both attempted to avoid NEAR/SQUARE tokens in their interactions with the participants, although as these initial interactions were not recorded, we cannot be sure of the degree to which this was successful.

In our analysis of the production data, we noticed that participants who interacted with the US experimenter produced a significantly greater distinction in the minimal pairs lists than those who interacted with the New Zealander, and they also reported more word pairs as being distinct. For example, for those who interacted with the US experimenter, 65% of all word pairs were reported to be distinct in their own speech, as opposed to 41% for those who interacted with the NZ experimenter. This tendency, together with similar experimenter effects we have seen in other recent experiments at Canterbury, led us to include the identity of the experimenter as an independent variable in our analysis of the perception data.

We now turn to the results of the perception task.

5. Results

In order to interpret the results of the perception task, we applied two separate statistical models of the perception data. Both models involve logistic regression models, which were fit using Harrell's Design Library in R (Harrell, 2004; The R Development Core Team, 2005).

The first model was a logistic regression analysis of the entire data set, including the degree to which the presence or absence of a photo affected the likelihood of a participant making an error. The second was a logistic regression which excluded the condition in which the photo was absent, and considered the effects of specific aspects of the photos on participants' performance. We first discuss the model in which the overall predictors of the data are explored, without considering specific aspects of the photo.

5.1. Overall results for entire data set

In fitting the overall model, we considered a large number of potential predictors. These include:

(1) *Information about the participants*, including their age, sex, social class, degree of distinction in production (as measured by their Pillai score in List 1), and whether or not they report the word pair to be the same in their own speech (in List 3).

(2) *Information about the items*, including the log frequency of the target (as assessed by frequency of occurrence in the Wellington Corpora of Spoken and Written English Bauer, 1993; Holmes, Vine, & Johnson, 1998), the log frequency of the competitor (i.e. the 'incorrect' word choice which appeared on the screen), whether the item was from the NEAR lexical set or the SQUARE lexical set, the degree to which the word pair was kept distinct by all participants (as measured by the word pair Pillai score from List 1, calculated across all participants).

(3) *Experiment/contextual effects*, including how far through the experiment the word was presented, whether the individual was in group 1, or in a photo condition, and the actual Euclidean distance between the specific word pair tokens being responded to (i.e. the distance between F1 and F2 during the first element of the respective vowels as assessed by: $\sqrt{(F1_a - F1_b)^2 + (F2_a - F2_b)^2}$.[5] Also included was information about the identity of the researcher. As noted above, two researchers were responsible for running the experiment—one who spoke with a New Zealand accent and another who spoke with an American English accent. We also considered aspects of the stimulus voice—including the identity of the voice as a factor, as well as the Pillai score for the voice, and the perceived age and social class of the voice. Obviously with only four voices, the danger of overfitting this aspect of the data set is rather large, and these variables could not all be tested together.

[5]We use the Euclidean distance rather than the Pillai here because we are comparing two tokens, rather than two distributions.

Table 7
Overall perception model: Wald Statistics for predicting error

	χ^2	*d.f.*	*P*
Participant Pillai (Factor + Higher order factors)	139.86	2	<0.0001
All interactions	11.47	1	0.0007
Interviewer (Factor + Higher order factors)	24.73	2	<0.0001
All interactions	11.47	1	0.0007
Photo presence	4.18	1	0.0408
Stimulus Euclidean distance	29.62	2	<0.0001
Nonlinear	15.86	1	0.0001
Near/square (Factor + Higher order factors)	61.03	5	<0.0001
All interactions	38.22	4	<0.0001
Item Pillai (Factor + Higher order factors)	18.68	2	0.0001
All interactions	8.70	1	0.0032
Position in experiment	8.54	1	0.0035
Self-report as "same"	6.27	1	0.0123
Participant age	9.30	1	0.0023
Log freq. (Factor + Higher order factors)	29.29	4	<0.0001
All interactions	28.98	3	<0.0001
Log competitor freq. (Factor + Higher order factors)	26.44	4	<0.0001
All interactions	23.90	3	<0.0001
Participant sex	7.33	1	0.0068
Participant Pillai × interviewer (Factor + Higher order factors)	11.47	1	0.0007
Near/square × item Pillai (Factor + Higher order factors)	8.70	1	0.0032
Log freq. × log competitor freq. (Factor + Higher order factors)	18.33	2	0.0001
Near/square × log freq. (Factor + Higher order factors)	26.75	2	<0.0001
Near/square × log competitor freq. (Factor + Higher order factors)	22.49	2	<0.0001
Near/square × log freq. × log comp. freq. (Factor + Higher order factors)	16.34	1	0.0001
Total interaction	51.24	6	<0.0001
Total nonlinear+interaction	59.07	7	<0.0001
Total	364.88	19	<0.0001

A logistic regression model was fit by hand, starting with a (relatively) saturated model, and removing non-significant factors. As indicated above, a fully saturated model with the above combination of factors was not possible, so some fitting was done by substituting comparable factors (e.g. to establish whether the Pillai score for the voice was a better predictor of error than the perceived social class of the voice). Not all possible interactions between all of the above factors have been tested for—but we did test for all two- and three-way interactions which seemed to us to be plausible. No four-way interactions were tested.

The ANOVA table for the overall fitted model is given in Table 7, and the model coefficients are given in Table 8. A large number of these factors proved highly significant. The significant effects are discussed below, grouped into different types of effects—participant-specific effects, item effects, and experimental/contextual effects.

5.1.1. *Participant effects*

Males misidentified words at a greater rate in this task than female participants, and the error rate decreased with increasing age.

How distinct a speaker tends to keep the two word pairs themselves (as measured by their Pillai score in production List 1) was significant. Unsurprisingly, the less merged the participants were themselves, the more accurately they performed on the task.

On top of this effect was an extra effect of whether the participant reported the specific word pair under consideration to be the 'same' or 'different' in their own speech. If the participant answered 'same', they were more likely to make an error for that item pair in the perception task.

Table 8
Overall perception model for predicting error—model coefficients

Absent model	
Intercept	−1.521904884
Participant Pillai	−1.072881978
Interviewer = US	0.687497442
Photo = present	−0.183201160
Stimulus Euclidean distance	0.310582342
Stimulus Euclidean distance (nonlinear)	−0.898743432
Near/square = NEAR	4.039004571
Item Pillai	−6.277194266
Position in experiment	−0.004860101
Self-report = same	0.208130908
Participant age	−0.013678021
Log freq.	0.284677087
Log competitor freq.	0.422545540
Participant sex = male	0.221159504
Participant Pillai × interviewer = US	−1.484019973
Near/square = NEAR × item Pillai	5.960893365
Log freq. × log competitor freq.	−0.093141724
Near/square = NEAR × log freq.	−1.015678577
Near/square = NEAR × log competitor	−1.412342626
Near/square = NEAR × log freq. × log competitor	0.274208806

5.1.2. *Item effects*

SQUARE words had significantly higher error rates than NEAR words. This is predicted by our previous model (Warren et al., forthcoming), which argues that a prelexical processor is biased toward mishearing infrequent items as more frequent items, leading to a general SQUARE as NEAR mishearing bias. This effect is predicted to be strictly prelexical.

The degree to which the word pair involved tends to be kept distinct by the population (as estimated by the item Pillai score, calculated for each item pair over all participants' productions in List 1) was a strong predictor of participants' performance. That is, the more distinct a word pair tends to be kept in individuals' accumulated experience, the more accurate they will be at distinguishing that word pair in speech perception. This also replicates a correlation reported in Warren et al. (forthcoming).

Unexpectedly, we found an interaction between whether the word is a NEAR or SQUARE word, and the item Pillai score. Namely, the more merged a word pair is, the larger the bias toward mishearing SQUARE (see Fig. 5). In fact, for word pairs which tend to be kept relatively distinct, there is no such bias at all. This is not predicted by our earlier position (Warren et al., forthcoming), where we attribute the SQUARE as NEAR bias to a prelexical effect. This interaction seems to provide strong evidence that the bias stems (at least partially) from the lexicon itself. Further problematic for our previous position is the fact that the item Pillai score is, in fact, not predictive of the NEAR errors (which were predicted to be largely lexical). While merged items have many more SQUARE errors than non-merged items, there is little difference in the NEAR error rate across the various degrees of merger. This interaction is highly problematic, and suggests that aspects of Warren et al.'s model will need some reformulation.

Also intriguing is a three way interaction between whether the item is a NEAR word or a SQUARE word, the log lexical frequency of the target word, and the log lexical frequency of its competitor (i.e. the frequencies of the two words that participants were choosing between on the screen).

Fig. 6 shows the relationship between the frequency of the target word and whether that word is a NEAR word or a SQUARE word. This graph is shown for three different values of the frequency of the competitor—the 1st quartile, median, and 3rd quartile.

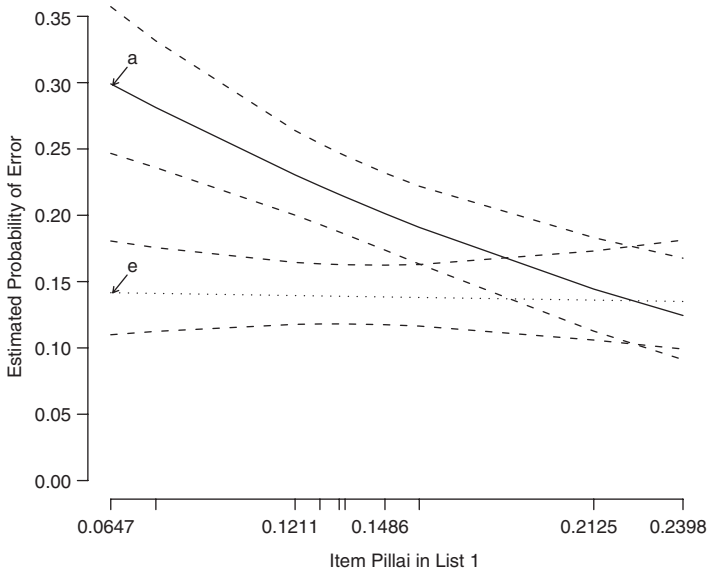

Fig. 5. Estimated probability of error for NEAR ("e") and SQUARE ("a") target words, as a function of how distinct that word pair is kept. Lower pillai scores indicate merged word pairs, higher pillai scores indicate word pairs where a greater distinction is maintained. The dashed lines indicate 95% confidence intervals.

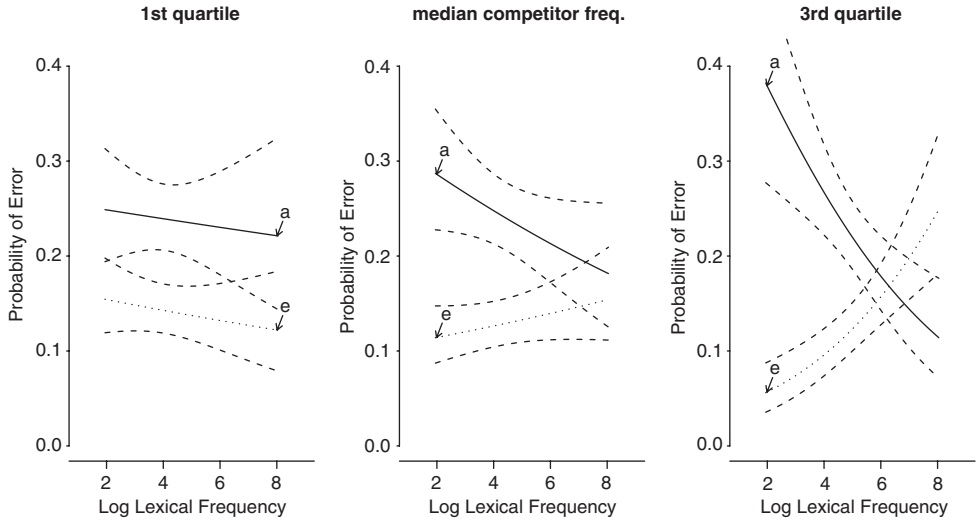

Fig. 6. Estimated probability of error as a function of target frequency, competitor frequency and NEAR/SQUARE items. Separate lines are shown for NEAR ("e") targets (dotted line) and SQUARE ("a") targets (solid line). The *x*-axis shows the frequency of the target word. The three panels show three different values of the competitor frequency: the 1st quartile, median, and third quartile. The dashed lines show 95% confidence intervals.

These graphs show that SQUARE errors tend to decrease with the increasing lexical frequency of the target word. The competitor exaggerates this pattern—such that if both words are high frequency, the SQUARE errors are particularly low, but if the target is low frequency and the competitor is high frequency, the SQUARE errors are particularly high.

For NEAR, if the competitor frequency is low, then NEAR errors decrease with the increasing frequency of the target word. However, as the competitor frequency increases, this relationship changes. If the competitor frequency is high, NEAR errors increase with increasing frequency of the target word.

While considering the graphs in Fig. 6, it is worth remembering that the experiment contains just 10 minimal pairs. Some parts of the target frequency/competitor frequency plane are populated with real observations, but these are quite sparse, relative to the continuous space depicted in Fig. 6. Nonetheless, the interaction is certainly significant, and seems to be intriguing. We return to further discussion of this below.

5.1.3. *Experimental/contextual effects*

Whether or not the photo was present was a significant predictor of participants' responses. The presence of a photo significantly decreased participants' likelihood of error. Recall that the photo appeared before the stimulus was played, and the voices were not blocked. Within any single condition, specific photos were uniquely paired with specific voices. Thus, in addition to any effect of social information, the photo also served simply as a cue as to which voice to expect next. It is not particularly surprising that in a perception task involving multiple speakers, providing a cue as to the upcoming voice will significantly aid perception.

The degree to which the specific word pair involved was kept distinct by the stimulus speaker played a role. While we made an attempt to select words which were roughly comparable in their degree of separation, there was inevitably some variation. Our speakers produced some word pairs more distinctly than others, and this significantly influenced participants' accuracy in the task. This influence was nonlinear, with low error rates for words with large Euclidean distances, and highly variable error rates (as evidenced by the broad confidence interval) for stimuli with smaller Euclidean distances (Fig. 7, left panel).

Participants got better at the task as the experiment progressed—the later a stimulus was presented in the experiment, the lower the error rate (Fig. 7, right panel).

Intriguingly, the identity of the experimenter played a role, with participants interacting with the United States (Californian) researcher showing a higher overall error rate in the experiment. This effect interacted with the Pillai score of the participants (Fig. 8). As noted above, participants with a lower Pillai score had more errors overall. In addition, the error rate for participants with lower Pillai scores is particularly high for those who met with the US researcher. For those with the highest Pillai scores, the US researcher if anything increased the accuracy slightly. In sum, the overall effect of the US researcher was to increase the error rate for those participants who are themselves relatively merged.

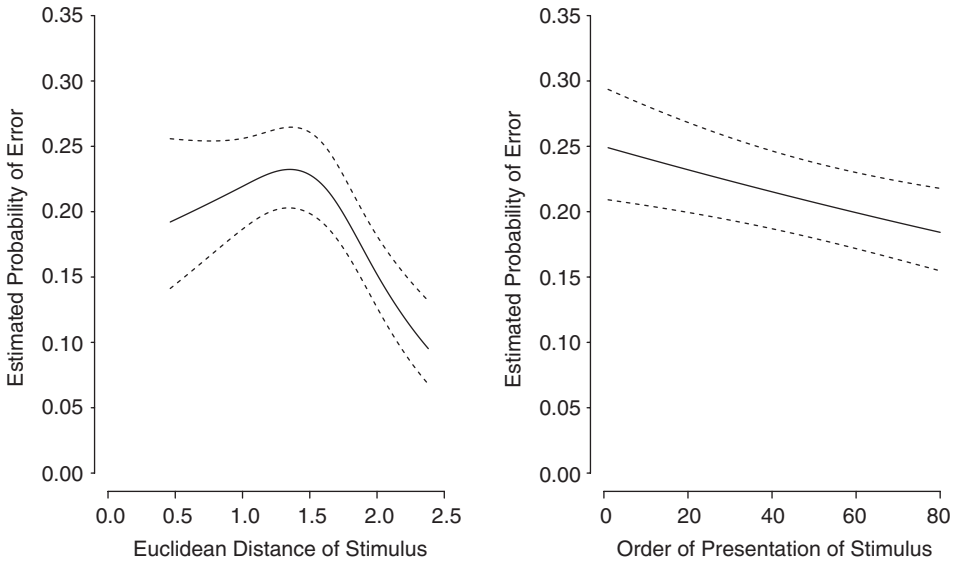

Fig. 7. Left panel: Euclidean distance of stimulus; right panel: order of presentation of stimulus. Dashed lines show 95% confidence intervals.

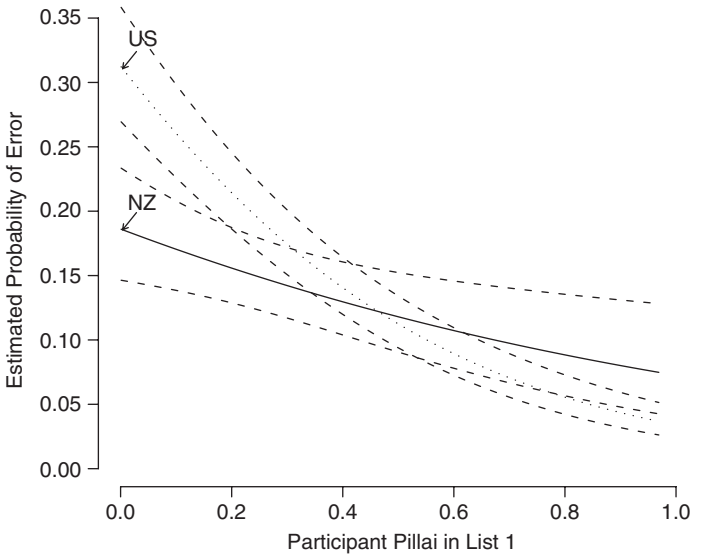

Fig. 8. The effect of the interviewer on error rate, depending on the degree of distinction maintained by a participant (as measured by their pillai score). Dashed lines show 95% confidence intervals.

5.2. Interim discussion

5.2.1. Effect of the interviewer

We had observed that in production, the US researcher tended to lead to an increased maintenance of the distinction. In perception, intriguingly, the US researcher actually increased error rates, at least for those speakers who are themselves merged. How can it be that interacting with a US researcher could lead individuals to more easily produce a distinction between NEAR and SQUARE, but reduce their accuracy at perceiving a distinction? This is particularly curious given the strong relationship we otherwise find between production and perception. In general, the more distinct an individual is in production, the more accurately they perform in the perception task.

In order to unravel this, we need to consider the fact that, while a speaker of Standard American English will make a distinction between NEAR and SQUARE lexical items, the phonetic realizations of these items will be quite different from those made by a NZE speaker who maintains the distinction. In particular, American English SQUARE has a much lower first element than either NEAR or SQUARE in NZ English.

The key to understanding the perception results comes from the observation that no NZ SQUARE variant is as low as the US variants—including the experimental stimuli. Both NZ SQUARE and NZ NEAR are closer to US NEAR than they are to US SQUARE. Thus, the more activated US variants are, the more everything produced by the stimulus speakers will resemble the NEAR distribution rather than the SQUARE distribution.

While non-merged participants are likely to have sufficient distinct NZ exemplars to perform the task, merged participants, not finding enough evidence in their local exemplars, may more readily look to the recently activated US exemplars in order to help them with their task. However, the US exemplars will be phonetically dissimilar to the variants produced by the NZE speakers, and so will lead to an overall higher error rate.

5.2.2. Item effects: lexical vs. prelexical

As outlined above, participants in the experiment displayed a strong bias toward responding with the NEAR lexical item, resulting in a high error rate for SQUARE. This replicates our past findings on this merger (Warren et al., forthcoming). Because participants will have encountered many SQUARE words with NEAR realizations, but few NEAR words with SQUARE realizations, this bias appears difficult to explain at a lexical level. In Warren et al. (forthcoming) we account for this bias by positing a fast phonological pre-processor which is biased toward mishearing infrequent items (SQUARE) as frequent items (NEAR). Thus, we predict this SQUARE as NEAR

bias to be entirely prelexical. However several results in experiments reported here cast considerable doubt on this interpretation.

First, the SQUARE bias interacted with the item Pillai scores. Words which tend to be more merged display a much higher SQUARE error rate than NEAR error rate. Words which are more often kept distinct show no such bias (Fig. 5).

Moreover, while the SQUARE errors are strongly affected by the degree to which a word pair tends to be kept distinct, the NEAR errors show no such pattern. This is exactly the opposite from what our earlier position would predict. Merged items contain close exemplars in both the NEAR and SQUARE lexical item, and so the more merged an item is, the more evidence that a NEAR stimulus might actually represent a SQUARE word.

We have been extremely puzzled by this interaction, which is highly statistically robust, and it has forced us to reconsider the source of the NEAR bias. We now believe that the best way to account for this bias is if participants' behavior is affected by the presence of non-NZE exemplars.

All New Zealanders are heavily exposed to other dialects of English, particularly through television and film. All New Zealanders, then, have been exposed to distinct tokens of NEAR and SQUARE words. Even speakers of NZE who have distinct NEAR and SQUARE forms have considerably closer realizations of SQUARE than any dialect they would be frequently exposed to, including Australian, British and American English.

Perhaps in this highly artificial perception task, participants seek to focus on parts of their remembered distributions which do not overlap. When a focus on local, familiar variants will provide enough information to make a distinction, then this is the basis on which the decision is made. When not, the focus moves to more extreme parts of the distribution, and evidence from other varieties of English is used.

Perhaps, then, for word pairs which New Zealanders are more likely to keep distinct, the participants rely on more central parts of their distributions—the error rate is in general lower, and there is no bias toward responding NEAR. However, for word pairs which tend to be merged, there is less evidence in this part of the distribution for the participants to reliably distinguish between NEAR and SQUARE. They then look to the more extreme parts of their remembered distributions for evidence of a distinction between the words. A bias toward responding NEAR emerges, as all of the stimuli more closely resemble the NEAR of other varieties than the SQUARE.

This interpretation may also provide some insight into understanding the interaction of NEAR and SQUARE lexical items with the lexical frequency of the target word and its competitor. The task involves comparison of an input with distributions of the potential candidates that are indicated by the words on the screen, i.e. a potential target and its competitor. If the frequency of either the target or the competitor is low in the NZE exemplars experienced, then broader experience is brought into play.

Fig. 6 showed the relationship between the frequency of the target word and whether that word is a NEAR word or a SQUARE word. Following the above discussion regarding the role of non-NZE exemplars, we might expect a relationship between the frequencies of the words, and the NEAR/SQUARE bias. Namely, the bias toward errors on SQUARE should be highest when the SQUARE word itself is infrequent. This lack of frequency may provide insufficient evidence on which the participants can make a decision, causing them to consult non-NZE exemplars. This will have the result of making the stimulus appear more NEAR-like.

When the target is SQUARE (labelled "a" on the graph), we predict the error rate to decrease with the increasing frequency of the SQUARE word. Lack of frequency leads to consultation of non-NZE exemplars. In Fig. 6 we see declining rates of SQUARE errors with increasing target frequency, consistent with this interpretation. We also see increasing rates of SQUARE errors with increasing competitor frequency, presumably due to a general bias to mishear less frequent things as more frequent things.

When the target is NEAR (labelled "e" on the graph) errors are predicted when participants consult NZE exemplars. This is because [iə] items will occupy both NEAR and SQUARE distributions. The more non-NZE exemplars are consulted, the more the NEAR target will resemble the NEAR distribution (relative to the competitor SQUARE distribution), and the fewer errors there should be. Thus, for NEAR items, we predict an increase in error rate with increasing frequency. The more frequent the item is, the more participants will be able to rely on NZE exemplars. Thus, in Fig. 6, we see NEAR errors increasing with the frequency of the target word.

When the target is NEAR, a low frequency SQUARE would lead to consultation of non-NZE SQUARE exemplars, decreasing the degree to which the target resembles the SQUARE distribution. Thus, we predict the more frequent SQUARE is, the higher the error rate for NEAR targets will be. For high frequency SQUARE, there is a lack of contrasting non-local exemplars, and there will be a general bias toward hearing less frequent things as more frequent things. This is true for the higher values of NEAR target frequency. For these values, error rates increase with the increasing frequency of the square competitor.

Only one thing remains mysterious about Fig. 6, and that is the low rate of NEAR errors predicted when NEAR is low frequency and SQUARE is high frequency. This is slightly mystifying, as it would seem that a lower frequency SQUARE should decrease errors—both because there is no frequency bias toward answering SQUARE, and because of the potential presence of contrasting non-NZE exemplars. We have no explanation for why the opposite is observed. For low NEAR target frequency values, a higher SQUARE frequency decreases errors. This is mysterious. However we note that this portion of the models' prediction is not based on real observations. That is, there are no word pairs in our data set where NEAR is very low frequency and SQUARE is very high frequency. The largest SQUARE/NEAR ratio is between *spare* (log frequency = 4.2) and *spear* (2.48). This particular part of the model, then, is based on interpolation, and—while intriguing—requires further work in order to substantiate the trends.

The interaction between NEAR/SQUARE, lexical frequency and the frequency of the competitor provides severe challenges for an account in which a NEAR bias arises prelexically. However it provides some support for the claim that participants rely partially on non-NZE exemplars in this task—particularly when evidence from native exemplars is weak. This evidence may be weak either because the distributions are highly merged, or because the frequency of the target word is low.

Whether the NEAR bias is a sole function of the reliance on non-NZE exemplars, or whether the prelexical processor is also involved remains a topic for further investigation.

5.2.3. The presence of the photo

One overall effect of particular interest in this data set is the effect of the presence of the photo. In this initial analysis, we looked to see whether the presence of a photo was having any effect. We found that participants performed significantly better when a photo was present. We regarded this as promising evidence that social information in the photo may be influencing participants' performance. However this is by no means the only possible interpretation. One likely effect of the photo is simply to cue the participant to the upcoming voice. The experiment included four voices, and these were not blocked. Thus, the association of a single photo with each voice for each participant then, served as a cue as to which voice to expect next. It is likely that this effect could be achieved without photos—simply by associating a unique symbol with each voice and flashing it on the screen before that voice were played.

5.3. Results of photo manipulation

Having established the general non-photographic effects responsible for variation in the data set, as well as the fact that having a photo present reduces the overall error rate, we then turned to an investigation of how, if at all, the specific photos affect participants' performance. To do this, we refit the statistical model described above to the subset of the data set for which a photo was present (i.e. excluding group 1).

As can be seen by the ratings in Tables 3 and 4, while we attempted to control for perceived social class in groups 2 & 3, and age in groups 4 & 5, this had limited success. The 'age-manipulation' photos differed substantially in perceived social class, and in the 'class-manipulation' photos, the same individuals differed in perceived age when photographed across the two class conditions. Rather than treat these as separate manipulations, then, we decided to fit a logistic model to all of the combined data from groups 2–5, including the perceived age and social class factors of each photo as potential predictors. Tables 3 and 4 show the 'occupation' and 'education' factors to be highly related. We therefore added these two scores together for each photo to create a combined 'social class' variable.

We first fit the previously described overall model (Table 7) to the subset of the data for which a photo was present (excluding, of course, the 'photo present' variable). All predictors remained significant. We then

Table 9
Model incorporating photo effects: Wald Statistics for photo-related variables

	χ^2	*d.f.*	*P*
Photo class (Factor + Higher order factors)	16.85	3	0.0008
All interactions	15.93	2	0.0003
Participant Pillai (Factor + Higher order factors)	101.24	3	<0.0001
All interactions	18.73	2	0.0001
Voice Pillai (Factor + Higher order factors)	7.76	2	0.0206
All interactions	7.15	1	0.0075
Self-report as "same" (Factor + Higher order factors)	12.39	2	0.0020
All interactions	6.76	1	0.0093
Photo age (Factor + Higher order factors)	7.78	2	0.0204
All interactions	6.76	1	0.0093
Participant sex (Factor + Higher order factors)	15.13	2	0.0005
All interactions	8.64	1	0.0033
Voice sex (Factor + Higher order factors)	9.00	2	0.0111
All interactions	8.64	1	0.0033
Photo class × participant Pillai (Factor + Higher order factors)	8.56	1	0.0034
Photo class × voice Pillai (Factor + Higher order factors)	7.15	1	0.0075
Self-report × photo age (Factor + Higher order factors)	6.76	1	0.0093
Participant sex × voice sex (Factor + Higher order factors)	8.64	1	0.0033
Total interaction	66.18	10	<0.0001
Total nonlinear+interaction	69.36	11	<0.0001
Total	285.04	26	<0.0001

Table 10
Model incorporating photo effects: coefficients of photo-related variables

Present model	
Photo class	4.066172149
Participant Pillai	−5.154802488
Voice Pillai	29.897448135
Photo age	−0.017737318
Participant sex = male	0.529191449
Voice sex = male	0.111111123
Photo class × participant Pillai	0.567936471
Photo class × voice Pillai	−4.634506280
Self-report = same × photo age	0.021407320
Participant sex = male × voice sex = male	−0.581032243

attempted to add in the photo-specific variables. We tested for interactions between these variables and other factors, and also re-tested for the effect of the sex of the speaker—hypothesizing that the presence of the (clearly gendered) photo may aid the emergence of any speaker-sex effect.

Three further factors emerged as significant, all in interaction with other variables. Tables 9 and 10 show the ANOVA table and the co-efficients for these factors, as well as the factors they interact with. Probability levels for the factors not involved in these interactions remained comparable to the model reported in Table 7.

The first effect was an interaction between the sex of the participant and the sex of the speaker, as shown in Fig. 9. This interaction was not significant when considered over the entire data set, suggesting that the presence of the photo, which cues the sex of the upcoming voice, is largely responsible for this effect. While male voices were responded to (marginally) more accurately by male participants than female participants, the interaction is carried almost entirely by the female voices, which were responded to significantly more accurately by the females than the males.

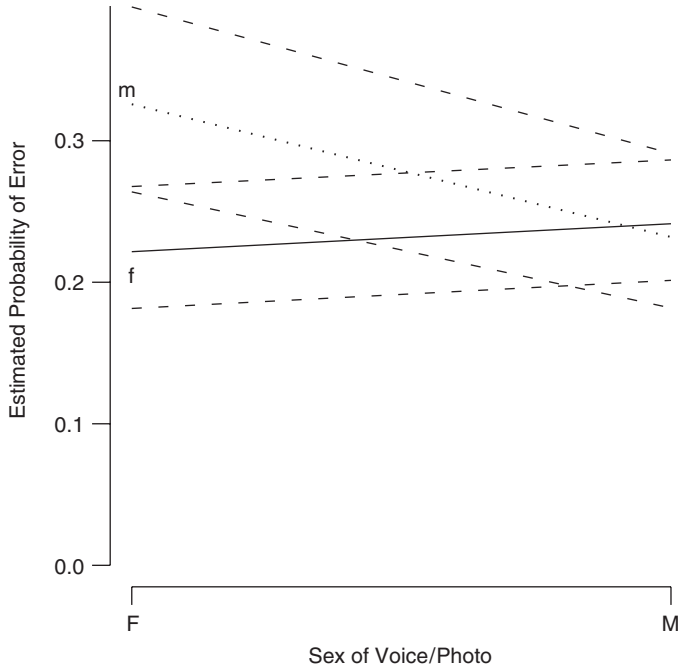

Fig. 9. The estimated probability of error for male (m) and female (f) participants when listening to male (M) and female (F) voices. Dashed lines show 95% confidence intervals.

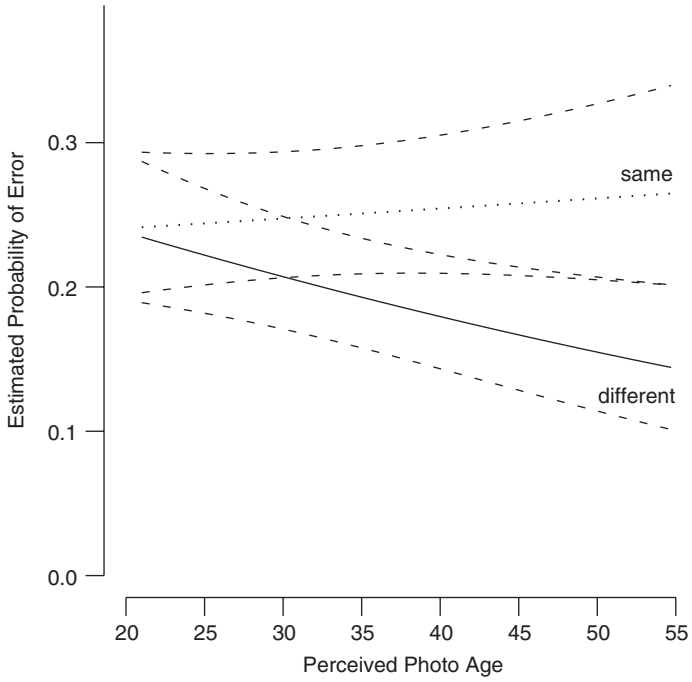

Fig. 10. The influence of the photo age on the probability of error if the word-pair was reported as "same" or "different". Dashed lines show 95% confidence intervals.

The second effect was an interaction between the perceived age of the photo and whether the participant self-reported the specific word pair as being the same in their own speech. This is shown in Fig. 10. Participants who reported the word pair as merged were not particularly sensitive to the age of the photo. If

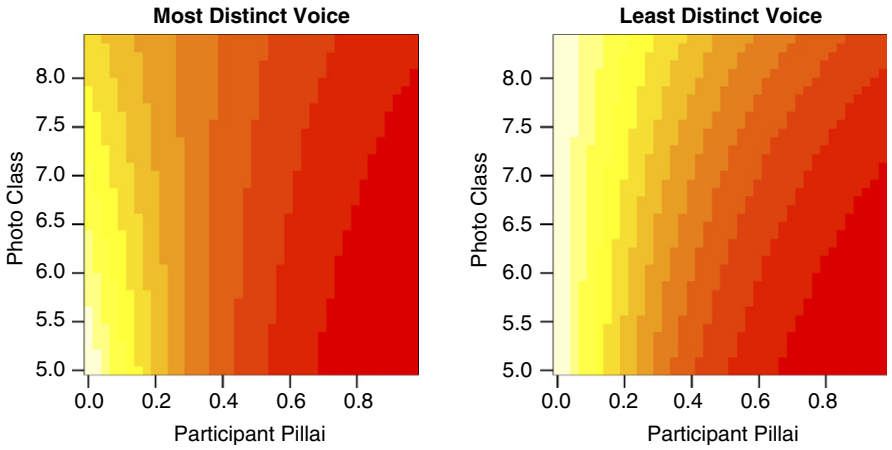

Fig. 11. The effect of perceived photo class on error rate, for different values of participant pillai, for two different voices. Lighter colors indicate higher error rates.

anything, an older photo increased errors for these speakers. However, participants who reported the word pair as distinct made significantly more errors with younger looking photos. For the youngest photos, those who reported a word pair to be distinct were only marginally more accurate than those who reported the word pair to be the same. However for the older photos, there is a considerable difference. This suggests that people who themselves make a distinction can be differently sensitive to the phonemes depending on the perceived age of the speaker. However the sensitivity of people who themselves do not make a distinction is not affected by perceived age.

Finally, the perceived social class of the person in the photo played a role, and this interacted with the participants' own Pillai score, as well as the degree to which the voice kept the two distributions distinct.

Fig. 11 shows these relationships. Each graph shows the interaction between the participant Pillai score and the perceived social class of the photo. The leftmost graph shows this interaction for the most distinct voice, and the rightmost graph shows the interaction for the least distinct voice. The graphs are image plots, in which lighter colors correspond to higher error rates. For both figures, the right side of the graph is darker, as we would expect, because the overall error rate of non-merged (high-Pillai) speakers is lower than the error rate of merged speakers. The graph for the least distinct voice contains more lighter color in general, reflecting the overall higher error rate.

For the most distinct voice, intriguingly, for the participants who don't make a distinction, the error rate appears to decrease with decreasing social class. However for those who do make a distinction, the social class of the photo does not have a strong effect.

This effect looks different for the least distinct voice, where there is little effect of the photo for the merged people, but the non-merged people react to the photo in the opposite way from that predicted—namely they make more errors when the voice is paired with a higher social class photo.

This different behavior of the voices is shown more directly in Fig. 12, which shows the relationship between the photo's perceived social class and the voice Pillai. The figure shows the model's predictions for these variables when the Participant Pillai is set at 0. The pattern is most clear for merged participants, as they were the most sensitive to the social class manipulation (cf. Fig. 11).

The voices which maintain the largest distinction between NEAR and SQUARE (the two with the larger Pillai scores) show an increase in errors with the decreasing social class of the photo. However, for the least distinct voice, the error rate actually slightly decreases with the decreasing social class of the photo. It is important to note that even the least distinct voice was quite distinct, with a Pillai score of 0.863. Of the 71 participants in the experiment, only 3 produced a distinction with a greater Pillai score than this.

Thus, both the age of the photo and the social class of the photo seem to have an effect on participant accuracy. The direction of the effect with age is exactly as predicted. The nature of the effect with social class is

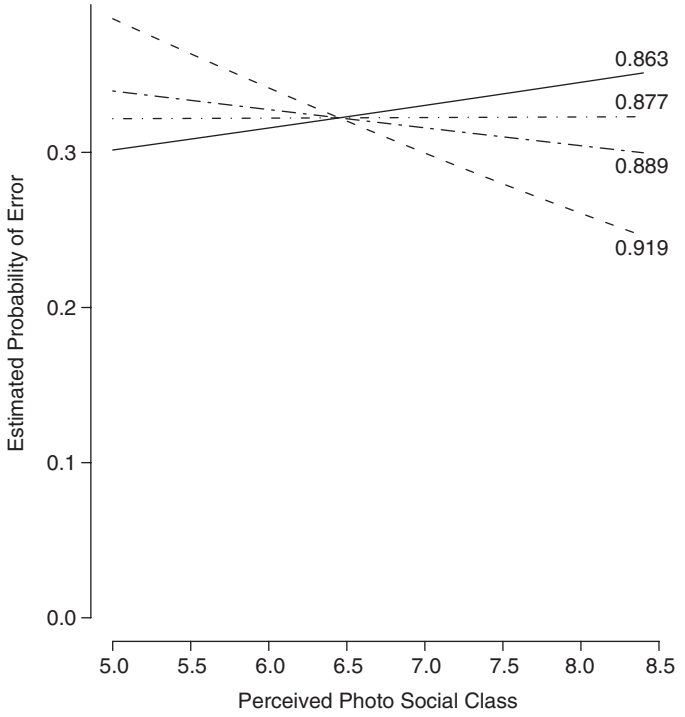

Fig. 12. Estimated probability of error for different values of perceived photo social class, and different stimulus voice pillai scores.

more complex, and interacts with how distinct the participant keeps the phonemes, and also how distinct the stimulus voice keeps them.

5.4. Discussion: effect of the photos

The fact that social aspects of the photos affect participants' performance provides strong evidence that individuals are sensitive to social information in speech perception. We regard this as support for our hypothesis that exemplars are indexed with social information.

Perceived age does not affect individuals who report that they do not make a distinction themselves. Their errors are uniformly high, regardless of the age of the person in the photo. However, those who report that they do make a distinction show relatively sophisticated sensitivity to social factors. When they believe they are listening to an older speaker, they attend more to the distinction, and perform more accurately on the task. When they believe they are listening to a younger speaker, they appear to treat the words as more ambiguous, and their accuracy is considerably reduced. This is despite the fact that the auditory stimuli are identical.

The effect of perceived social class is more mysterious. As predicted, the social class of the photos affects participants' accuracy in this task. However the direction of the effect is not entirely as predicted. For the two most distinct voices, errors increase with the decreasing social class of the photo. For the least distinct voice, errors increase with the increasing social class of the photo.

This seems to indicate the importance of compatibility between the perceived class of the photo, and the degree of distinction maintained by the voice. Incompatibility between the perceived class of the photo (which presumably sets up an expectation about degree of distinction) and the actual distinction maintained leads to higher error rates. The fewest errors occur with the most distinct voice when paired with the highest social class photo.

On top of this is an effect of the degree to which the participant maintains the distinction themselves. Those who are relatively merged display the predicted behavior. Their error rates decrease with the increasing social

class of the photo. As predicted, a working class photo leads to activation of exemplars which are relatively more merged than a middle class photo.

What we did not predict was that those who maintain a distinction themselves would show the opposite trend. They are, in fact, more likely to make errors with middle class than lower class photos. This is particularly surprising because it was the non-merged participants who showed the expected sensitivity to photo-age.

We believe the difference between perceived age and perceived class for this group stems from the fact that while almost everyone is exposed to speech from a variety of ages, not everyone is exposed to speech from a variety of social classes. Lower social classes will be exposed to the speech of higher social classes via the media. But the reverse is not necessarily true, and one can certainly imagine that individuals might not interact with sufficient individuals of lower socio-economic status than themselves to have a sizeable inventory of socially indexed 'lower class' exemplars. Thus it is entirely possible that listening to a younger speaker activates younger exemplars, thus decreasing sensitivity to the distinction, whereas speech from a lower class speaker does not provide a sufficient set of appropriately indexed exemplars to perform the task. If there are insufficient exemplars, activation will spread through the remainder of the distribution. Because the speaker makes a distinction themselves, this should provide a solid basis on which to make an accurate response.

This interpretation is strengthened by results recently reported by Drager (2005b), who found a similar interaction in her study of the perceived boundary between two vowels undergoing chain-shift in NZE. She found an interaction between the social class of a photo associated with a stimulus voice and the social class of participants. Crucially, this interaction was evident not only in participants' responses, but also in their reaction times. That is, working class participants were faster to respond to voices associated with working class photos than middle class photos, and the opposite was true for middle social class participants. The effect of the photo on reaction time was markedly stronger for the middle class participants than the lower class participants. If a group is familiar to you, the raised activation level of socially appropriate exemplars will speed overall performance. However if there are few such exemplars, recognition will need to rely more on the overall acoustic match of the input to the candidate distributions—with activation spreading until the candidate distributions are sufficiently distinct to make a decision.

6. General discussion

Our results clearly show that social information affects speech processing. The exact mechanism through which this occurs is not so clear, and there is much work to be done in further investigating the nature of the effects identified here. We believe that the most successful model of our results is very likely to be an exemplar model, in which encountered speech is stored as exemplars, which are indexed for both linguistic and social information. Establishing the precise details of such a model will require much further investigation, as is well demonstrated by the discussion provided by other contributors to this volume. Here, we simply offer our best attempt at understanding the source of the combined effects reported in this paper. We suggest that the following factors are involved in influencing participants' behavior in our task:

(A) Participants come to our task with differing experience, and so differing sets of exemplars. In addition, their existing exemplars will have different degrees of resting activation, depending (in part) on what speakers the participant has recently interacted with.

(B) Participants rapidly detect that the experiment involves particular vowels (or a particular vowel). The presentation of each photo then raises the activation level of socially relevant exemplars containing the relevant vowel(s). That is—exemplars are activated in the NEAR and SQUARE lexical sets (for unmerged participants), or in the NEAR/SQUARE combined lexical set (for merged participants).

(C) Presentation of an auditory stimulus raises the activation level of acoustically similar exemplars.

(D) Visual presentation of the two words between which participants have to choose causes participants to consult activation in the two candidate distributions, and respond with the candidate which is most activated.

(E) If there is not enough evidence to make a decision based on (D), activation spreads beyond the initially activated exemplars until there is enough evidence. Or, if there is never enough evidence, the participant guesses.

The differing starting points of the participants (A) is responsible for the participant-specific effects we observe—including the effect of gender and age on participants' performance in this task, as well as the effect of the participants' own production of the vowels. (A) is also the locus of the effect of the researcher. Exemplars indexed as US English receive slightly elevated activation for those participants who interact with a US experimenter.

(B) involves activation of exemplars with the appropriate social indexing. We argue that this activation may involve any exemplars containing NEAR and SQUARE vowels. When the photo is first encountered, participants do not know which minimal pair to expect. The potential candidates do not appear on the screen until after the word is played. Thus, the raising of the activation level of socially appropriate exemplars in (B) must certainly include exemplars beyond the minimal pair involved. Thus we suggest that (B) involves elevation of activation for socially appropriate exemplars which are indexed as belonging to the appropriate lexical set (or containing the appropriate 'phoneme'). The development of exemplar models of speech perception has involved discussion of distributions of both phonemes (e.g. Johnson, 1997; Pierrehumbert, 2001b) and words (e.g. Bybee, 2001). Pierrehumbert (2003, p. 178) argues that in viable theories there is a "ladder of abstraction, each level having its own representational apparatus". We thus assume an exemplar approach, in which encountered words are stored as complete phonetic memories, and are indexed to appropriate categories such as 'male', 'female', (Johnson, 2006, Foulkes and Docherty, 2006), as well as to appropriate linguistic categories such as phoneme labels. This phonemic indexing will presumably differ across individuals with different systems. In particular, participants who have sufficiently distinct representations for NEAR and SQUARE exemplars presumably index these separately, such that *hear*, *fear*, *beer* etc. share a label, and *hare*, *fare* and *bare*, share a different label. Participants who do not maintain the distinction index all such words as containing vowels which belong to the same category.

This difference in phonemic indexing is responsible for the differential effect of perceived age on different participants. We assume that unmerged participants will anticipate the upcoming stimuli by raising the resting activation levels of all NEAR/SQUARE indexed items produced by speakers with the appropriate social characteristics. Because NEAR and SQUARE items will be indexed separately, this will activate relatively distinct distributions for older voices, enabling a distinction to be readily heard. However activating NEAR/SQUARE labelled items produced by younger voices will lead to activation of less distinct distributions. Because the younger photo activates NEAR and SQUARE categories which are in fact, largely identical, performance degrades to near the level of the merged participants.

For merged participants, while exemplars for, e.g. *here* and *hair* occupy separate distributions, both are indexed to the same, collapsed, phonemic category. This is regardless of the age of the voice. Thus, when the older photo is encountered, the activation of 'older' tokens indexed to this collapsed category does nothing to facilitate activation of distinct distributions.

Note, however, that the merged participants can still do very well at this task (despite the fact that many report they are guessing). This is because, despite the identical phonemic labelling of the lexical items, the phonetic memories still occupy distinct exemplar clouds—that is, they still occupy different word-level distributions. Thus, while (B) does nothing to help them with the task, at (C) the acoustic signal will still match the appropriate exemplar cloud better than it matches the competitor. This will generally lead even merged participants to get the answer right more often than they get it wrong, even though the absence of overt 'phonemic' discrimination leads them to feel that they are guessing. (C), then, involves activation of word-level distributions.

The cumulative social, phonemic and lexical activation will, in many cases, lead to participants having a clear answer when the two choices are presented to them (D). If there are exemplars which are activated both by the social priming and acoustic input then these will clearly dominate, and are likely to lead to a clear decision. However error rates increase when social information and acoustic information do not match. This was exemplified in the interaction (in Fig. 12), between the degree of distinction maintained by a voice and the social class of the photo with which it was associated. The least distinct voice increased in errors with a higher social class photo, but the most distinct voice decreased in errors with the higher social class photos.

Activation may spread beyond the set of initially activated exemplars, especially if this set of exemplars is inadequate for a decision (E). For example, for our non-merged participants, we hypothesized that their exemplars of lower-class speech may be extremely sparse. Thus, when presented with a lower class photo, the

socially relevant exemplar set was inadequate to reach a decision. The reference sets therefore expanded to include a larger number of distinct tokens, and participants performed the task relatively accurately.

One way in which the candidate sets may be inadequate to reach a decision is if the sets for the two candidate words are non-distinct. The sets may then expand until they include non-local exemplars, introducing a NEAR bias (because, even for distinct speakers, the NZ SQUARE resembles other varieties' NEAR). This bias therefore emerges for word pairs which are highly merged, but not for word pairs which tend to be kept relatively distinct.

Another way in which the initial evidence may be inadequate is if one or other of the candidate items is relatively infrequent. Activation may then spread to include non-local exemplars. If the target item is SQUARE, low frequency will increase the overall error rate (because non-local exemplars provide misleading information). If the target item is NEAR, low frequency may slightly decrease the overall error rate (because non-local exemplars provide helpful information).

Participants who make the distinction themselves are likely to have relatively distinct exemplar sets, and so are able to perform the task more successfully than participants who do not make the distinction, and whose exemplar sets are likely to be more overlapping. The latter participants are apparently influenced by the presence of a US English speaking researcher, which facilitates activation of the more distinct, non-New Zealand exemplars.

Note that we are not suggesting that the continuing expansion of the candidate set occurs in a conscious or explicit manner. A distribution is activated first in the most socially/demographically/regionally/acoustically relevant part of the distribution, and then activation spreads to more peripheral exemplars, For sparse distributions, more marginal exemplars will be activated more quickly than for dense distributions. And when comparing highly overlapping distributions, a decision may be delayed until sufficiently distinct distributions have been activated, no matter how peripheral.

This interpretation (together with results reported by Drager, 2005b) suggests that timed tasks may be particularly revealing. The results presented here may have looked quite different if participants were required to respond in a very short time frame.

The multi-faceted nature of these effects also points to the potential benefits of considering exemplar accounts of speech perception alongside those which have developed in different domains. For example exemplar accounts of social judgement and stereotype formation argue that memories of individual people are stored in memory as exemplars, and are retrieved in the process of forming social judgements (e.g. Smith & Zarate, 1992). These exemplars can be accessed from memory, in order to influence judgements, and this occurs independently of whether the individual is consciously aware of the prior experience. There are clear points of connection here, pointing to the potential benefits of a careful investigation of the ways in which individuals gradually built up categories that emerge from the experiences that surround them as social actors. There is certainly good evidence that these categories are both linguistic and social, and that individual exemplars may be simultaneously encoded along both dimensions.

In addition to the theoretical implications of these results, the large number of social factors which apparently play a role in speech perception should cause experimenters some nervousness regarding experimental methodology. Whenever the production of a target variable displays social variation, the social characteristics of the participants, as well as the social attributes they attribute to the stimuli are likely to play a role in participants' behavior in perception tasks. This suggests that there will be many experimental designs in which researchers will need to carefully consider the social distribution of their participants, and relevant aspects of their experimental stimuli. Even more worrying is the apparent effect of recent interlocutors. It is obviously difficult to completely control who our participants have recently interacted with. Minimally, we should aim to ensure that the same researcher runs all of the participants in a particular experiment.

7. Conclusion

We set out to test the hypothesis that exemplars are socially indexed, and that individuals are sensitive to social information in speech perception. The results of our experiment were relatively complex. While many of the specific interpretations we offer for the variety of results reported here will need further experimental investigation, we believe the general hypothesis receives strong support. We have struggled to explain the

complex interactions between participant-specific characteristics, word-specific characteristics, context-specific characteristics, and perceived speaker characteristics. But this struggle would have been infinitely greater outside an exemplar based approach.

The fact that error rates are highest for words which tend to be most merged in the participants' experience strongly supports an exemplar approach. Moreover, there is good evidence that the exemplars are socially indexed. The perceived social characteristics of an unknown speaker affect how their speech is processed. Individuals appear to attempt to match the stimulus to the relevant subparts of the candidate distributions. Where these subparts are not adequate to make a decision, activation spreads to less relevant parts of the same distribution. Activation of less central parts of the distribution is facilitated by exposure to a speaker of a dialect which maintains a clear distinction.

This study has revealed sound-change to be a rich area for the study of spoken word recognition. Studies which link production and perception data in the context of sound-change in progress promise to reveal much about the speech perception system, and how it deals with a system which is in a state of flux.

Acknowledgements

We are grateful to those people who donated their voices and faces for experimental stimuli, to Brynmor Thomas and Alice Murphy for their help with data collection and analysis, to Harald Baayen for statistical advice, and to Christian Langstrof and Margaret Maclagan for providing comments on an earlier draft. This paper has also greatly benefited from the comments of Stefanie Jannedy, and three anonymous reviewers. Support for this research was provided by a University of Canterbury research grant to the first author, and a University of Canterbury summer scholarship to the third author.

References

Batterham, M. (2002). The apparent merger of the front centring diphthongs—EAR and AIR—in New Zealand English. In A. Bell, & K. Kuiper (Eds.), *New Zealand English*. Wellington: Victoria University Press.

Bauer, L. (1993). *Manual of information to accompany the Wellington corpus of written New Zealand English* (pp. iv + 131). Wellington: Victoria University, Department of Linguistics.

Bayard, D. (1987). Class and change in New Zealand English: A summary report. *Te Reo*, *30*, 3–36.

Bayard, D., Weatherall, A., Gallois, C., & Pittam, J. (2001). Pax Americana?: Accent attitudinal evaluations in New Zealand, Australia, and America. *Journal of Sociolinguistics*, *5/1*, 22–49.

Bybee, J. (2001). *Phonology and language use*. Cambridge: Cambridge University Press.

Davis, P., Jenkin, G., & Coope, P. (2003). *NZSEI-96*: *An update and revision of the New Zealand socio-economic index of occupational status*, Statistics New Zealand, Wellington.

Davis, P., McLeod, K., Ransom et al. (1997). *The New Zealand socioeconomic index of occupational status* (*NZSEI*): *Research Report #2*, Statistics New Zealand, Wellington.

Di Paolo, M., & Faber, A. (1990). Phonation differences and the phonetic content of the tense-lax contrast in Utah English. *Language Variation and Change*, *2*, 155–204.

Drager, K. (2005a). From Bad to Bed: an investigation of vowel perception in New Zealand English. *Te Reo*, to appear.

Drager, K. (2005b). *The influence of social characteristics on speech perception*. Masters Thesis, University of Canterbury, unpublished.

Elley, W. B., & Irving, J. C. (1985). The Elley–Irving socio-economic index: 1981 census revision. *New Zealand Journal of Educational Studies*, *20*, 115–128.

Foulkes, P., & Docherty, G. (2006). The social life of phonetics and phonology. *Journal of Phonetics*, *34*(4).

Fry, D. B. (1947). The frequency of occurrence of speech sounds in Southern English. *Achives Néerlandaises de Phonétique Expérimentale*, *XX*, 103–106.

Gimson, A. C. (1963). *An introduction to the pronunciation of English*. London: Edward Arnold.

Gordon, M. J. (2002). Investigating Chain Shifts and Mergers. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *Handbook of language variation and change* (pp. 244–266). MA: Blackwell Publishers Ltd.

Gordon, E., & Maclagan, M. (2001). Capturing a sound change: A real time study over 15 years of the NEAR/SQUARE diphthong merger in New Zealand English. *Australian Journal of Linguistics*, *21*(2), 215–238.

Harrell F. E (2004). Hmisc S function library. Programs available from 〈http://biostat.mc.vanderbilt.edu/s/Hmisc〉.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*, 373–405.

Hay, J., & Maclagan, M. (2002). Does spelling influence perception when sounds are merging? Poster presented at New Ways of Analysing Variation 31, Stanford, October 2002.

Hay, J., Pierrehumbert, J., & Beckman, M. (2003). Speech perception, wellformedness and the statistics of the lexicon. In J. Local, R. Ogden, & R. Temple (Eds.), *Papers in laboratory phonology VI: Phonetic interpretation* (pp. 58–74). Cambridge, UK: Cambridge University Press.

Holmes, J., & Bell, A. (1992). On shear markets and sharing sheep: The merger of EAR and AIR diphthongs in New Zealand English. *Language Variation and Change*, *4*, 251–273.

Holmes, J., Bell, A., & Boyce, M. (1991). Variation and change in New Zealand English: A social dialect investigation. *Project Report to the Social Sciences Committee of the Foundation for Research, Science and Technology*, Victoria University, Wellington.

Holmes, J., Vine, B., & Johnson, G. (1998). *Guide to the Wellington corpus of spoken New Zealand English*. Victoria University of Wellington, Wellington: School of Linguistics and Applied Language Studies.

Johnson, K. (1997). Speech perception without speaker normalization. In K. A. Johnson, & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–166). San Diego, CA: Academic Press.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, *34*(4).

Johnson, K., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, *24*(4), 359–384.

Labov, W. (1994). *Principles of linguistic change: Internal factors*. Oxford: Blackwell.

Maclagan, M., & Gordon, E. (1996). Out of the AIR and into the EAR: Another view of the New Zealand diphthong merger. *Language Variation and Change*, *8*, 125–147.

Newman, R. S., Sawusch, J. R., & Luce, P. A. (2000). Underspecification and phoneme frequency in speech perception. In M. B. Broe, & J. B. Pierrehumbert (Eds.), *Papers in laboratory phonology V: Acquisition and the lexicon* (pp. 298–311). Cambridge, UK: Cambridge University Press.

Ohala, J. J. (1992). What's cognitive, what's not, in sound change. In G. Hellermann (Ed.), *Diachrony within synchrony* (pp. 309–355). Frankfurt: Peter Verlag.

Olson, C. L. (1976). On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, *83*, 579–586.

Pierrehumbert, J. (2001a). Why phonological constraints are so coarse-grained. *Language and Cognitive Processes*, *16*, 691–698.

Pierrehumbert, J. (2001b). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee, & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 137–157). Amsterdam: John Benjamins.

Pierrehumbert, J. (2002). Word-specific phonetics. In C. Gussenhoven, & N. Warner (Eds.), *Laboratory phonology*, Vol. 7 (pp. 101–139). Berlin & New York: Mouton de Gruyter.

Pierrehumbert, J. (2003). Probabilistic phonology: Discrimination and robustness. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 177–228). Cambridge, MA: The MIT Press.

Pitt, M., & Johnson, K. (2003). Using pronunciation data as a starting point in modelling word recognition. *Paper presented at the 15th international congress of phonetic sciences*, Barcelona.

R Development Core Team (2005). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria 〈http://www.R-project.org〉.

Savin, H. B. (1963). Word-frequency effect and errors in the perception of speech. *Journal of the Acoustical Society of America*, *35*, 200–206.

Smith, E. R., & Zarate, M. A. (1992). Exemplar-based model of social judgment. *Psychological Review*, *99*(1), 3–21.

Strand, E. A. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology*, *18*(1), 86–99.

Warren, P. (2005). Word recognition and word merger. In J. Luchjenbroers (Ed.), *Cognitive linguistic investigations across languages, fields, and philosophical boundaries*. Amsterdam: John Benjamins.

Warren, P., & Hay, J. (2005). Using sound change to explore the mental lexicon. In C. Fletcher-Flinn, & G. Haberman (Eds.), *Cognition, language, and development: Perspectives from New Zealand*. Bowen Hills, Queensland: Australian Academic Press.

Warren, P., Hay, J., & Thomas, B. (forthcoming). The loci of sound change effects in recognition and perception. In J. I. Hualde, & J. Cole (Eds.), *Laboratory phonology* (Vol. 9).

Warren, P., Rae, M., & Hay, J. (2003). Word recognition and sound merger: the case of the front-centering diphthongs in NZ English. In M.-J. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th international congress of phonetic sciences* (pp. 2989–2992). Rundle Mall: Causal Productions.

Watson, C., Harrington, J., & Evans, Z. (1998). An acoustic comparison between New Zealand, and Australian English vowels. *Australian Journal of Linguistics*, *18*(2), 185–207.

Wells, J. C. (1982). *Accents of English*. Cambridge: Cambridge University Press.