Available online at www.sciencedirect.com
ScienceDirect
Journal of Phonetics 37 (2009) 276–296
Journal of Phonetics
www.elsevier.com/locate/phonetics

Subphonemic and cross-phonemic priming in vowel shadowing: Evidence for the involvement of exemplars in production

Sam Tilsen

University of California, Department of Linguistics, 1203 Dwinelle Hall, Berkeley, CA 94720-2650, USA

Received 26 March 2008; received in revised form 15 February 2009; accepted 7 March 2009

Abstract

Despite strong evidence that cognitive representations of speech targets rely upon a mapping between perceptual and motor memories, the nature of those representations—whether they are stored exclusively as abstract categories or can incorporate more detailed episodic memory—remains an open question. A primed vowel-shadowing experiment was conducted to investigate the extent to which the recent perception of subphonemic details of vowel quality can influence subsequent productions. If such effects are observed, they argue for an exemplar-based model of production in which the mapping between perceptual and articulatory representations can occur rapidly and can incorporate subphonemic detail. On experimental trials, subjects were primed with one vowel before they heard a second vowel, which they shadowed. On some trials, the formants of the prime were subtly manipulated. Significant subphonemic priming effects were observed in the F1 and F2 of responses. In addition, cross-phonemic priming tended to be dissimilatory, which may be related to an analogous phenomenon observed in studies of manual and oculomotor movement control. Accounts of these findings are discussed in the context of exemplar models of speech perception and production.

1. Introduction

A current trend in linguistic research is a movement away from theoretical approaches that aim to generalize, accompanied by a movement toward theories and models which target variation and complexity. This movement is not restricted to linguistics, but more generally is associated with a new generation of cognitive science, one which recognizes the enormous mnemonic power of the brain, and which considers cognitive capacities that extend beyond a purely symbolic view of memory and its influence on behavior. In phonetics and phonology, evidence of this trend is manifested in recent special issue of the Journal of Phonetics (2006, *34*, 4) on modeling sociophonetic variation, and also in a recent special issue of *The Linguistic Review* (2006, *23*, 3) on exemplar-based models in linguistics.

From a cognitive perspective, any complete model of the complex variation associated with speech requires a satisfactory understanding of what sort of information is remembered about speech percepts, how that information is stored in memory, and in turn, how that memory affects the perception and production of speech. In line with the general movement toward understanding the complexity of cognitive systems, a number of speech researchers have recently designed models of speech perception and production that accommodate the power of the brain to remember very specific details about events. Goldinger (1996, 1998) and Johnson (1997) presented exemplar-based models of speech perception along these lines. In these approaches, linguistic variation is remembered in the form of distributions of variables, which are updated by experiences of individual percepts, each of which may be associated in memory with various linguistic and non-linguistic contextual information. Pierrehumbert (2001, 2002) extended this idea to a model of how exemplar-based memory may influence speech production. These approaches are well-suited for integration with larger scale models of language acquisition and diachronic change that involve interacting agents (e.g. Oudeyer, 2006; Pierrehumbert, 2004; Wedel, 2004), each with their own detailed memories. In sum, one effect of the trend, which is

E-mail address: tilsen@berkeley.edu

0095-4470/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wocn.2009.03.004

applicable to a range of linguistic subfields, is to shift the main topics of research from invariants and universality to variation and complexity.

The dominant alternative to understanding sociolinguistic variation, which is nicely described in Pierrehumbert (2006), has trouble accounting for the empirically evident range of variation in speech, because it necessarily treats memories as categorial. This alternative, sometimes called the "Varbrul program," is a way of thinking about the relation between sociolinguistic variables (such as gender, class, age, etc.) and allophonic variation. In the Varbrul program, allophonic variation (which is understood as symbolic variation) is probabilistically related to sociolinguistic variables, and suballophonic variation is simply not a concern. All the gradient variation in the speech output is then relegated to various anatomical, physiological, mechanical, and aerodynamic constraints, as was the case in Chomsky and Halle (1968). This essentially marginalizes fine-grained phonetic detail from the realm of cognition. Hence, the contrast with exemplar-based accounts could not be more striking: rather than putting suballophonic detail aside, exemplar approaches put a central focus on how such variation is remembered and influences production—this makes phonetic detail an object of cognitive science, and adopts a more cognitively realistic understanding of memory.

Although there are many advantages to approaches which give greater import to memory, these approaches raise a new problem. There are a number of types of memory commonly referred to by cognitive psychologists: long-term memory, short-term memory, working memory, episodic memory, procedural memory, implicit memory, declarative memory, semantic memory, etc. Moreover, there is overlap between many of these categories, and substantial disagreement about how distinct some of these categories are. For example, episodic memory can be understood as a set of details remembered in association with an experience, but it is appropriate to study such memory both as long-term memory, which may persist on ultradian timescales, and as working memory, which is recalled or activated transiently, often for some task-specific purpose. There is currently much disagreement over whether distinct cognitive systems give rise to long-term and working memory, or whether these types of memory rely primarily upon a common neural substrate (Raganath, Johnson, & D'Esposito, 2003; Suprenant & Neath, 2008). In other words, we can no longer assume that there exists a "verbal scratchpad" of phonological working memory which is functionally distinct from long-term phonological memory. Furthermore, the time course of perception-to-memory formation processes in not well-understood, particularly in speech. We should thus take care not to let these contested categories of memory unduly bias our understanding of speech processing. In the following, "episodic memory" will be used to refer both to the long-term representation of specific information and to the maintenance of such information in working memory.

1.1. Episodic memory in speech perception and production

An important issue in modeling speech perception and production is how episodic memory contributes to the representation of speech percepts and targets. Episodic (or exemplar-based) memories store various details associated with a specific event. In the case of a speech event, an episodic memory (or *exemplar*) includes information about the talker, location, time, associated emotions, etc., as well as detailed auditory traces of the words in the utterance. One key difference between episodic and other forms of memory is that episodic memories refer to specific events; in contrast, non-episodic memories are believed to arise from integration of multiple experiences—hence in an adult listener, exposure to a speech event creates in episodic memory new examples of the linguistic categories involved in the utterance, but only very slightly alters the non-episodic representations of those categories.

In this paper, we will be concerned foremost with the question of how episodic memories influence speech production. For these purposes, we will sidestep the issue of whether speech targets should be conceptualized primarily as perceptual or motor memories. Undoubtedly, both perceptual and motor domains play a role in the cognitive representation of a speech target, and any satisfactory theory of speech requires some form of a mapping between these two domains (cf. Liberman & Mattingly, 1985). With this agnostic stance, the issue of perceptual vs. motor representation of targets becomes more about *when* and *how* perceptual representations are mapped to motor representations.

Regarding *when*, models of speech production can differ with respect to whether the mapping occurs online or offline in the planning and production of speech movements. This difference corresponds to distinct predictions about whether very recent perceptions can influence productions. Fowler, Brown, Sabadini, and Weihing (2003) demonstrated that this mapping can occur online: in comparing response times in simple and choice shadowing tasks, they found relatively small (~25 ms) differences between the simple and choice conditions. This finding suggests that recent perceptions are mapped to gestures very quickly, because latencies in choice conditions—in which perceptual stimuli must be mapped to articulatory plans—do not greatly exceed those in simple conditions, where articulatory plans are prepared prior to the stimuli.

Regarding *how* the mapping occurs, the question we will address here is whether episodic memories (exemplars) or non-episodic perceptual representations are mapped to motor representations. In other words, do representations of speech targets incorporate subphonemic (and suballophonic) details of the sort that are retained in episodic memory, or are the targets derived exclusively from abstract, non-episodic memories? One can imagine that the perceptual-motor mapping involves the transformation of a relatively invariant and abstracted representation of a percept to similarly invariant gestural coordinates, in which

case very recent perceptions should not influence productions. Alternatively, if episodic perceptual representations are mapped to gestures, then very recent perceptions may exert observable influences on productions.

In the domain of perception, there exists ample evidence to support the view that perceptual representations of speech include episodic memories; a variety of perceptual phenomena are usefully understood as arising from episodic memory of speech events. For example, Hintzman, Block, and Inskeep (1972) found that voice details of isolated spoken words persist in memory. Goldinger, Pisoni, and Logan (1991) found improved recall for 10-speaker wordlists compared with 1-speaker wordlists presented at slow rates, which suggests that interspeaker voice variation is stored in long-term memory along with lexical information. Palmeri, Goldinger, and Pisoni (1993) found that phonetic details are retained for several minutes, and Goldinger (1996) found similar effects lasting up to a week.

Models of speech perception which utilize episodic memory for the representation of speech can account nicely for such observations. Goldinger (1998) showed he could account for his findings using the Hintzman (1986) MINERVA 2 model of episodic memory. The most relevant models for our purposes will be the speech perception model described by Johnson (1997, 2006), and the production model described by Pierrehumbert (2001, 2002), in which an exemplar is a set of associations between auditory properties (the output of the peripheral auditory system, including phonetic details) and category labels, which describe the speaker, setting, and various linguistic categories such as words or phonemes. Exactly what sort of information can constitute a category label and exactly how category labels arise is still not well-understood, although presumably multisensory integration and associative mechanisms are involved in their formation. In these approaches, a new item is classified by comparing it to previously stored exemplars: each exemplar becomes activated according to its similarity to the new item, and the summed activation of the exemplars in a given category constitutes evidence the new item belongs to that category. Each exemplar has a base level of activation, which decays over time. Consequently, more recent exemplars contribute more than older exemplars to the categorization of new percepts.

Exemplar approaches have several nice aspects not shared by non-episodic approaches. For one, an exemplar model obviates the need for talker normalization algorithms in perception. Because exemplar representations include category labels referencing age, gender, and other social variables, the process of perception is biased to perceive speech inputs that reflect the presence of those categories in a given context. Johnson (2006) shows that cross-linguistic variation in the effect of gender on vowel formants cannot be accounted for by hard-wired normalization algorithms, while an exemplar-based model can handle such variation. Exemplar storage can account for word-frequency effects in lexical recognition (Goldinger, 1998) and phonetic categorization (Ganong, 1980); assuming that the base activation level of an exemplar decays over time (memories decay), more frequent words will have higher summed activation levels because they are associated with more exemplars—hence all other things being equal, a new item is more likely to be classified as a member of a more frequent category.

In the domain of production, there are a variety of effects that suggest an influence of episodic memory. For one, subphonemic influences on productions were found in the shadowing task conducted by Fowler et al. (2003). In their experiment, there were two conditions, a simple task in which listeners shadowed the first V of a model VCV sequence and then produced a predetermined CV, and a choice task in which listeners shadowed the entire model VCV sequence. In addition to finding only a relatively small response time difference between the simple and choice tasks, the researchers found adjustments in VOT toward the model VOT—hence a subphonemic detail of the percept was found to influence production on a very short timescale. This suggests that an episodic representation of the model VCV was quickly mapped to a motor representation that incorporated subphonemic information. Goldinger (1998) also presented evidence for subphonemic influences on production in shadowing wordlists produced by a variety of talkers.

Word-specific phonetic patterns argue for including episodic memory in production models too. For example, Yaegor-Dror and Kemp (1992) and Yaegor-Dror (1996) showed that in Montreal French certain semantically or culturally defined groups of words display idiosyncratic vowel qualities, having failed to undergo an otherwise regular shift. As Pierrehumbert (2001) points out, for any production model to account for this sort of pattern, production targets must reference the specific lexical items they are associated with in a given utterance—the sort of association that can be attributed to episodic memory.

Pierrehumbert (2001, 2002) presented an exemplar model of production that can account for such findings. In this model, perception occurs basically as described in Johnson (1997): new items are categorized according to their similarity to existing items in an exemplar space, which includes dimensions for each phonetic value that the perceptual system stores in memory (exactly what phonetic values are involved in this process remains an open question). A production target is determined by randomly selecting one exemplar from a subgroup of exemplars, and then taking an activation-weighted average of nearby exemplars. The activation strength is a gradient value which influences the probability an exemplar will be chosen for a production goal. Because the activations of exemplars decay over time, more recent exemplars are more likely to be selected. Due to the weighted averaging, more recent and frequent exemplars also contribute more to the determination of production targets.

2. Method

2.1. Primed vowel shadowing

To address whether recent perceptions affect speech production, this study employed a primed-shadowing task with two vowel phonemes, /a/ and /i/. On half of all trials, subjects first heard a prime (or "cue") vowel, and then after a short delay, heard a target vowel; on these trials, subjects repeated the target vowel as quickly as possible. However, on a quarter of the trials, the target stimulus was a pure tone beep ("no-target" trials), in which case subjects repeated the cue vowel (cf. Fig. 1). The no-target trials encouraged subjects to plan to produce the cue vowel to a greater extent than the non-cue vowel, because the cue vowel was twice as likely as the non-cue to be the required response. In the remaining control trials, subjects heard a beep for the cue stimulus, and then shadowed the target vowel. Note that the control condition is comparable to a two-choice vowel-shadowing task.

The relations between the cue and target stimuli, and their proportions, are schematized in Fig. 1. The lines between the cue and target stimuli represent the different types of trials. The vowels [a] and [i] were used for both target and cue stimuli; F1- and F2-centralized versions of these vowels, [a]* and [i]*, were used for the cue stimuli as well. The differences between the normal vowels and their centralized counterparts were subphonemic, approximately 50 and 70 Hz for F1 and F2, respectively.

On each trial, subjects heard a short stretch of white noise preceding the cue. Then they heard the cue stimulus, which was either [a], [i], [a]*, [i]*, or a BEEP. This was followed by an interstimulus delay of either 100 or 800 ms, which was balanced across conditions. After the delay, the target stimulus, which was either [a], [i], or a BEEP, was heard, and subjects responded as quickly as possible.

Trials in which the cue and target stimuli belong to the same phoneme will be called *concordant* trials, and those in which cue and target belong to different phonemes will be called *discordant* trials. The inclusion of no-target trials is crucial in this design, for the following reason: whenever the cue was a vowel, the probability of that same vowel being the required response was 2/3, and the probability of the other (non-cue) vowel being the required response was 1/3. This bias, along with a motivation to respond quickly, encouraged subjects to plan to say the cue vowel to a greater extent than the non-cue vowel; subjects also may have adopted a biased articulatory configuration prior to hearing the target—this possibility will be discussed in Section 4. If productions are influenced by phonemic or subphonemic details of the cue stimuli, then we should conclude that very recent perceptual experiences can exert effects on the formation of speech targets.

The exemplar model of production described in Pierrehumbert (2001, 2002) allows for recently perceived subphonemic details to influence the determination of production targets. This is because more recent exemplars are more highly active and thus more influential in the activation-weighted averaging of phonetic values through which production targets are determined. The comparison of primary interest in this experiment is between response vowel formants on normal-cue and centralized-cue

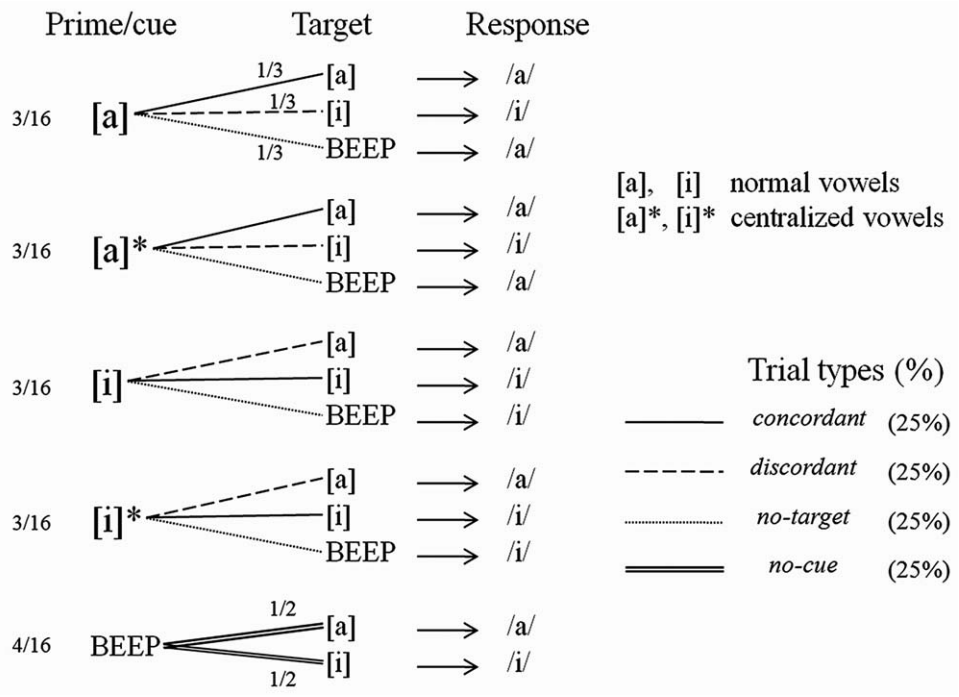

Fig. 1. Schematization of primed vowel-shadowing experimental design. Possible relations between prime and target stimuli are shown. Lines between prime and target stimuli indicate trial type: concordant trials (solid line), discordant trials (dashed line), no-target trials (dotted line), and no-cue trials (double line). Proportions of prime-target relations are shown, and percentages of trial types.

concordant trials, and is made explicit by the following hypothesis:

Hypothesis. *Rapid subphonemic perceptual-motor mapping*—on concordant trials, responses made after centrally shifted cue stimuli will be more central than those made after normal cues.

This hypothesis entails, for example, that [a] responses after centralized [a]* cues will tend to be more central in F1–F2 space than those made after normal [a] cues. This would indicate that the subphonemic differences in the cue stimuli were perceived and influenced the planning of vowel targets. We will in addition consider whether cross-phonemic priming effects occur, i.e. whether vowel responses will differ between concordant and discordant trials.

2.2. Experimental design

Subjects were 18–40-year-old native speakers of American English with no history of speech or hearing problems. 12 subjects participated, 6 males and 6 females. All subjects took part in 2 or 3 one-hour sessions, over the course of which they performed a total of 25–40 blocks of 32 trials, which amounts to around 800–1200 trials for each subject. Each trial began with an interval of white noise of random duration from 1000–4000 ms, followed by a 100 ms interval of silence. The white noise was intended to disrupt any residual effects from the preceding trial. The duration of the noise was randomized in order to avoid the establishment of a rhythm from the onset of the noise to the cue stimulus. After the 100 ms interval of silence, subjects heard the cue stimulus, which was either a beep or one of four vowels, two of which were /a/, the other two /i/. The F1 and F2 of one vowel from each phonemic category were shifted slightly to make the vowels more central in F1–F2 vowel space (cf. Section 2.3 for a description of how formant manipulation was accomplished). All stimuli were 250 ms in duration.

Following the cue was a delay of either 100 or 800 ms. These provide approximately 550 and 1250 ms intervals between the end of the cue stimulus and the typical response onset time. If response planning precedes response onset by approximately 100–200 ms, the delays provide about 400–1100 ms of time between the cue vowel and response planning. These durations were chosen to provide intervals of time differing in the relative extent to which planning of the cue stimulus might influence subsequent response planning and execution. After the interstimulus delay, subjects heard the target stimulus, which was either a beep or one of the two unshifted vowels, [a] and [i]. Note that if the cue stimulus was a beep, the target stimulus was restricted to a vowel, so that beep–beep trials never occurred (cf. Fig. 1). Following the target stimulus was a 2000 ms interval in which the subject's response was recorded.

Each trial can be characterized by three control parameters: cue stimulus, interstimulus delay, and target stimulus. Each block of trials consisted of 32 trials, 16 of which represented all permutations of the four cue vowels, two delay conditions, and two target vowels (these are the concordant and discordant trials, cf. Fig. 1). Eight trials consisted of a beep cue followed by the two target vowels in both delay conditions, all repeated twice in each block (no-cue/control trials). The remaining 8 trials consisted of the four cue vowels with a beep target in both delay conditions (no-target trials). Note that hearing any given cue vowel made that vowel twice as likely as the non-cue vowel to be the required response. This imbalance was expected to encourage greater planning of the cue vowel on all trials. The order of trials was randomized within each block to discourage subjects from guessing at the next response.

After each block of trials (except for the first two of each session), subjects received feedback regarding the speed of their responses in the block. This feedback came in the form of rating numbers which indicated how quickly they responded relative to their past response times in the session. The rating numbers were computed by using the means of the response times in the last completed block as arguments to the inverse cumulative normal distribution function, the parameters of which were estimated from the means of the response times in all prior blocks in the session. Thus the ratings ranged from 0 to 100, with values near 50 indicating that the average response time in the last block was close to the average mean response time in all preceding blocks. This system had the advantage that as subjects responded more quickly, it became more difficult to achieve higher ratings, and thus they had to maintain a high level of attention over the course of a session if they wanted to get high ratings. In order to facilitate concentration on the task, subjects were given a 5 min break halfway through each session.

2.3. Stimuli

Vowel stimuli were constructed with the following procedure: a speaker of Midwestern American English who makes no distinction between a low back vowel /ɑ/ and a mid back vowel /ɔ/ produced sets of approximately 100 tokens each of the vowels /ɑ/ and /i/. The vowel /ɑ/ is normally more front in most American English dialects and so will here be represented as /a/. The tokens closest to the mean F1 and F2 of each set were selected as base tokens (vowel formants were estimated using an LPC algorithm implemented in Matlab, cf. Section 2.4 for details). Using PSOLA resynthesis, the pitch contours of both vowels were changed to 105 Hz with a slightly falling contour using the formula: $F0 = 105-20t$, where t is the time in seconds from the onset of the vowel. The first 250 ms (over which the pitch fell from 105 to 100 Hz) of the signals were windowed using a Tukey window with $r = 0.25$ to reduce the salience of onset transients, and all stimuli were normalized to have the same signal energy.

Centralized versions of the stimuli were constructed using a method of formant shifting described in Purcell and Munhall (2006). The base tokens produced with the procedure described above were bandpass filtered in a narrow range of frequencies above or below the formant being shifted, and likewise the signals were bandstop filtered in a range of frequencies containing the center of the formant. The two filtered signals were resynthesized to produce a signal in which the manipulated formant was shifted in the direction of the bandpass filter. This method was chosen over LPC resynthesis because it produces more natural-sounding formant-shifted stimuli. De-emphasis was accomplished with a 3rd-order elliptical filter with 2 dB of passband ripple and 50 dB of stopband attenuation, and emphasis was accomplished with a 3rd-order elliptical filter with 0.5 dB of passband ripple and 15 dB of stopband attenuation. The filtered signals were resynthesized at a 1:1 ratio. Passbands and stopbands for F1 and F2 were, respectively (relative to the unshifted formants, in Hz), $F1_{pass}([a]) = [-150\ 0]$, $F1_{stop}([a]) = [0\ 150]$, $F2_{pass}([a]) = [50,\ 250]$, $F2_{stop}([a]) = [-250,\ 50]$, $F1_{pass}([i]) = [25,\ 350]$, $F1_{stop}([i]) = [-75,\ 25]$, $F2_{pass}([i]) = [-350,\ -50]$, and $F2_{stop}([i]) = [-50,\ 200]$.

The locations of the stimuli formant values (averaged over the middle third of each vowel) are shown in Fig. 2 in F1−F2 space. Fig. 3 shows spectrograms of the four stimuli. Looking carefully at the spectrograms, one can see that F1 and F2 are slightly closer for [i]* than [i], and slightly further apart for [a]* than [a]. The formants of [a] fall slightly; F2-[i] decreases across the vowel, while F1-[i] rises slightly. The mean LPC-estimated F1 and F2 of the four vowel stimuli were [a] (696, 1151), [a]* (651, 1218), [i] (284, 2223), and [i]* (341, 2150). The differences between the normal and centralized stimuli were thus (−45, 67) for [a] and (57, −73) for [i]. These differences are within the normal range of formant variation that one would expect of these vowels in casual speech, and were unlikely to have been consciously perceived by unsuspecting, untrained ears. The impression one gets in hearing these stimuli is that the difference between the centralized [i]* and normal [i] is slightly more difficult to hear than the difference between centralized [a]* and normal [a]. There was a remote possibility that the centralized [a]* could have been perceived as a low and central /ʌ/, but no subjects reported hearing a third vowel or responded so as to indicate such a perception, and moreover the task instructions encouraged them to think of their responses as either one of the two vowel phonemes /a/ and /i/.

2.4. Response measurements

The primary dependent variables in this experiment were F1 and F2. Secondary dependent variables were response time (or reaction time, i.e. RT) and response duration (Dur). Response time was defined as the time from the onset of the target stimulus to the onset of vocal fold vibration. The first block that each subject completed was not included in the analysis. Occasionally, subjects responded late (more than three standard deviations greater than their mean RT), responded early, or failed to respond. Such trials were excluded. Trials were also excluded if a response duration was too short (less than

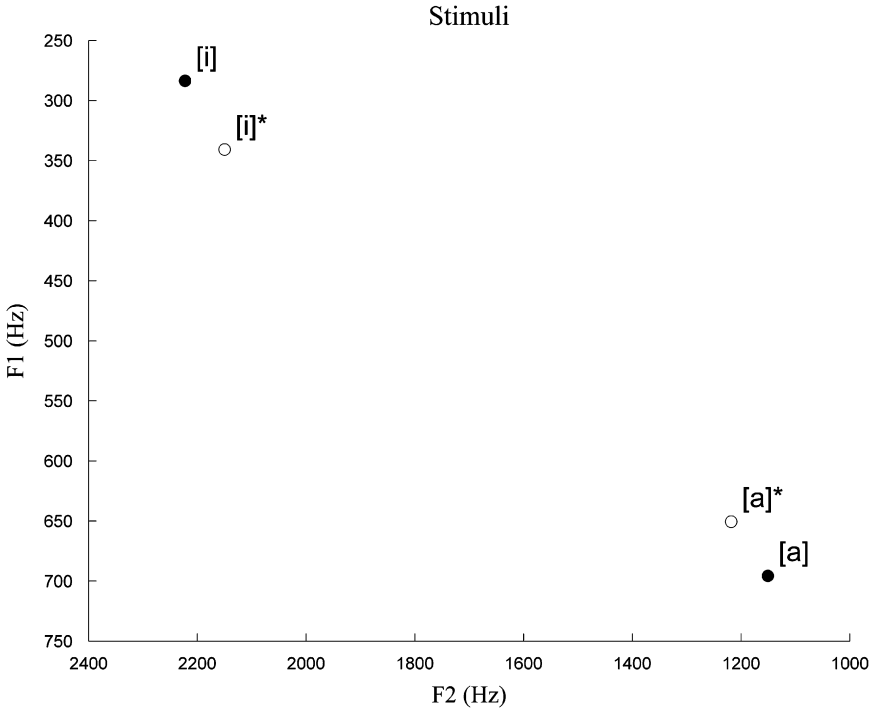

Fig. 2. Vowel space locations of normal and centralized stimuli (formants were averaged over the middle third of each vowel).

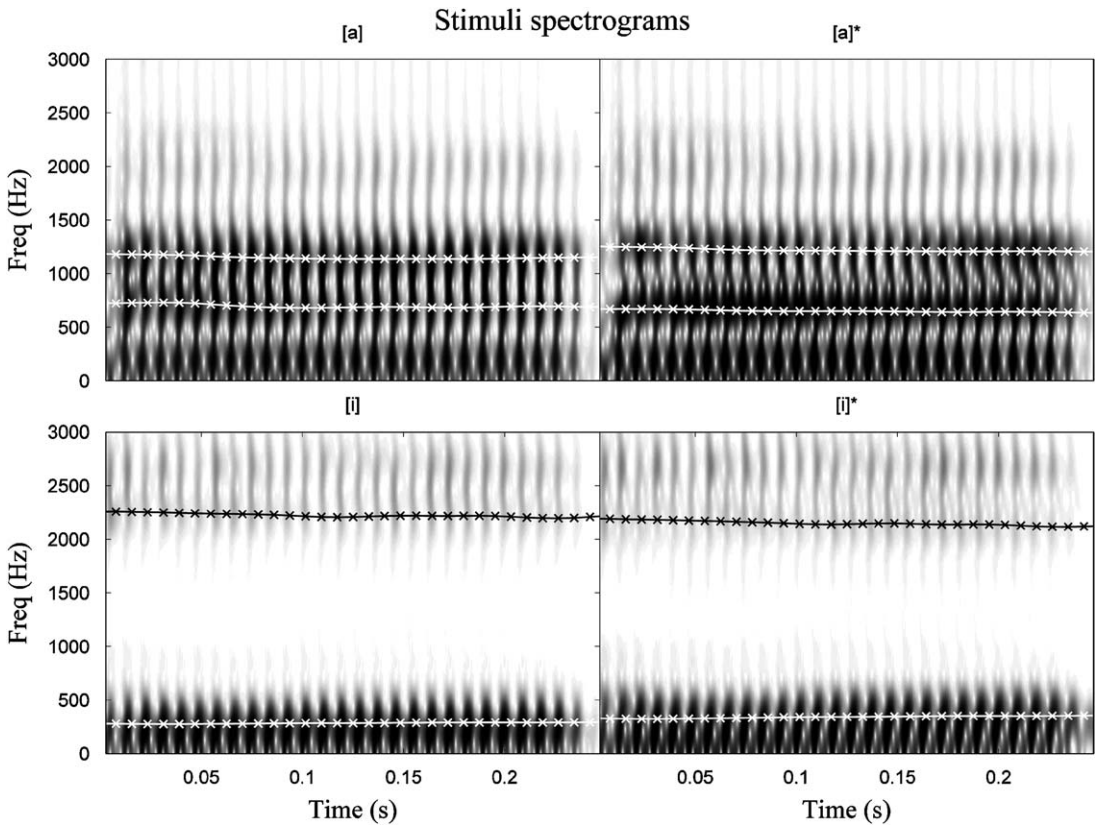

Fig. 3. Spectrograms of normal and centralized stimuli.

120 ms), the wrong response, mixed (i.e. a transition from [a] to [i] or vice versa), or had some other abnormality, such as being obfuscated by a non-speech vocalization. The mixed and incorrect responses occurred mostly on discordant trials.

Formants were estimated using an LPC algorithm implemented in Matlab. Responses were recorded at 44,100 Hz and downsampled to 11,025 Hz. 10 and 12 LPC coefficients were used for the male and female subjects, respectively; LPC coefficients were computed for 40 ms windows at steps of 5 ms. The first 10 ms of each response was skipped to avoid transient perturbations due to the onset of phonation. A pre-emphasized signal was used for measuring F1 and F2 of both responses, except for F1 of [i], where the pre-emphasis was found to occasionally interfere with the detection of a peak corresponding to F1.

Some further steps were taken to ensure robust formant measurement. When a formant for a given frame did not fall within a reasonable range, its value was interpolated from nearby formants; if reasonable formants were not found in ten consecutive frames or 12 frames total in the vowel, the LPC algorithm was considered to have failed and the trial was discarded. Finally, the formants were smoothed and averaged across frames. Most of the LPC algorithm failures ($<0.2\%$ of trials) involved low amplitude tokens in which the F1 and F2 spectral peaks of an [a] were indistinct or the F2 and F3 of an [i] were indistinct. Additionally, tokens with F1 or F2 values more than four standard deviations away from the mean for each subject were also discarded.

Overall, less than 5% of the entire dataset was excluded for the reasons mentioned above. When one discounts the late responses, less than 2% of all trials were excluded. Keeping the late responses makes no qualitative differences in the formant results presented in the following section. Because quick response times are indicative of attention to the task, and because lack of attention is possibly a confounding factor, it was judged best to exclude these RT outliers.

For the comparisons of mean formant values between conditions, means were averaged over the first, middle, and last third of each token. In the presentation of results below, the analysis of the mean formants from the middle third of each vowel is shown, unless otherwise stated. Analyses of variance (unbalanced, repeated measures) were conducted using type III sum of squares. Subjects were treated as fixed factors. Lack of balance in the data resulted primarily from differing numbers of observations between subjects. A very small imbalance within conditions was due to the excluded trials. Where contours are presented below, these contours were normalized for duration within subjects such that each response had the same number of LPC analysis frames as the mean number of frames.

3. Results and discussion

3.1. Subphonemic priming effects

Analyses of variance revealed significant mean F1 and F2 differences between response vowels produced after normal cues and centralized cues on concordant trials. This suggests that subphonemic details of the prime stimulus influenced vowel production targets. The effect of centralized cues on vowel response formants was significant for F2-[a] $\{F(1,1428) = 12.08,\ p < 0.001\}$, F1-[i] $\{F(1,1446) = 62.54,\ p < 0.001\}$, and F2-[i] $\{F(1,1446) = 6.04,\ p < 0.014\}$. Fig. 4 shows 95% confidence ellipses for the bivariate means on normal-cue, centralized-cue, and control (no-cue) trials, for each subject. These ellipses represent regions in which one can be 95% confident the true population mean vector is located. The tilt of the ellipses relative to the coordinate axes reflects the correlation between F1 and F2, and the lengths of the major and minor axes of the ellipses correspond to the variability of the samples in the directions of those axes. The figure shows that most subjects exhibit a tendency to produce relatively more centralized responses after a centralized-cue stimulus, particularly for the F2 of [a] and for both F1 and F2 of [i].

Several patterns can be seen in the relations between the normal, centralized-cue, and control trial means. Note that because there were twice as many control (no-cue) trials, their confidence regions are smaller. Where normal vs. centralized-cue differences were present, the most common pattern is for the control trial means to be relatively similar to the normal trials (e.g. [a]-*m*2; [i]-*f*1, *f*2, *f*3, *f*6, *m*1, *m*2, *m*3, *m*5). For some subjects, however, the control trial means are noticeably distinct from the normal and centralized-cue trials (e.g. [a]-*f*1, *m*1, *m*3, *m*5; [i]-*f*4, *f*5, *m*4, *m*6), although most of these remain more similar to the normal trials than the centralized-cue trials. However, for other subjects, the control trial means are located between the normal and centralized-cue trial means (e.g. [a]-*f*2, *f*5, *m*1, *m*6; [i]: *f*1, *f*5, *m*1, *m*2).

Within-subject comparisons of normal and centralized-cue trial formants were conducted using 1-sided *t*-tests (cf. the appendix for a table of within-subject statistics). 5 subjects exhibited individually significant differences for F2-[a], and all but two contributed to a trend for centralization. 9 subjects exhibited marginal or significant differences for F1-[i], and a clear trend can be seen across subjects to produce responses with higher F1 after shifted cue stimuli. Likewise, for F2-[i], 5 exhibited significant differences, and 9 followed the same trend.

A more temporal approach to the analysis of these data involves comparison of formant trajectories between normal and centralized-cue trials. Fig. 5 shows F1 and F2 contours for [a] and [i] responses made on normal, centralized-cue, and control trials. In general, the patterns of differences between normal and centralized-cue trials accord with the analyses of variance, but inspection of the trajectories allows for more detailed temporal patterns to emerge. First examining [a] responses (where lower F1 and higher F2 are more central), we see that the majority of subjects exhibited negligible differences in F1 throughout the contours, yet several subjects had relatively lower F1 in the first half of the contour (*f*3, *f*5, *f*6), suggesting centralization, while 2 subjects showed relatively higher F1 during some portions (*f*2, *m*2).

For [a]-F2, all male subjects tended to produce higher contours for part of or for the entire vowel, and several female subjects did so as well. One common pattern was for the centralized-cue trial formants to begin and remain central relative to the normal trials (*m*1, *m*3, *m*4). A different but no-less-prevalent pattern was for the trajectories to begin apart, then eventually converge (*f*1, *f*4, *m*5, *m*6). Another pattern is for the formants to begin near one another, but then the centralized-cue responses diverge centrally (*f*3), and in the case of (*f*5), subsequently reconverge.

For [i] responses, most of the male and female subjects clearly demonstrated centralization in F1, particularly during the first half of the response. As with [a], some subjects produced a relatively constant difference in formant value between the normal and centralized-cue conditions, while for others the formants converge eventually. The majority of subjects produced responses with more central F2, although a couple produced responses with more peripheral F2 (*f*4, *f*5).

In addition to the experimental trials (concordant), the no-target trials also offer a comparison between responses made after normal and centralized cues. The analysis of no-target trials revealed patterns of difference between mean formant values on normal and centralized-cue trials that were very similar to the patterns on concordant trials. The effect of centralization was significant for F2-[a] $\{F(1,1420) = 5.56,\ p < 0.019\}$, F1-[i] $\{F(1,1409) = 49.34,\ p < 0.001\}$, and F2-[i] $\{F(1,1409) = 10.14,\ p < 0.001\}$. These differences further support the hypothesis that subphonemic detail of the cue stimuli are incorporated into vowel target planning.

Centralization had no significant effects on response time. For [i] responses only, cue-centralization had a significant effect on response duration $\{F(1,1446) = 5.57,\ p < 0.018\}$, with responses tending to be longer after normal cues. Interstimulus delay did not exhibit any main effects on vowel formants on concordant trials, or any interaction effects with centralization.

The results presented above strongly support the hypothesis that subphonemic details of the cue stimuli would be perceived and integrated into response vowel planning. This suggests that episodic memories influence speech targets. This was true in particular for F2-[a], F1-[i], and F2-[i]. Although not every subject exhibited significant differences individually, the trends were highly significant across subjects. Because the same pattern held for comparisons between no-target trials, the effect does not depend upon the recency of the unconscious perception of a difference between a centralized cue and normal

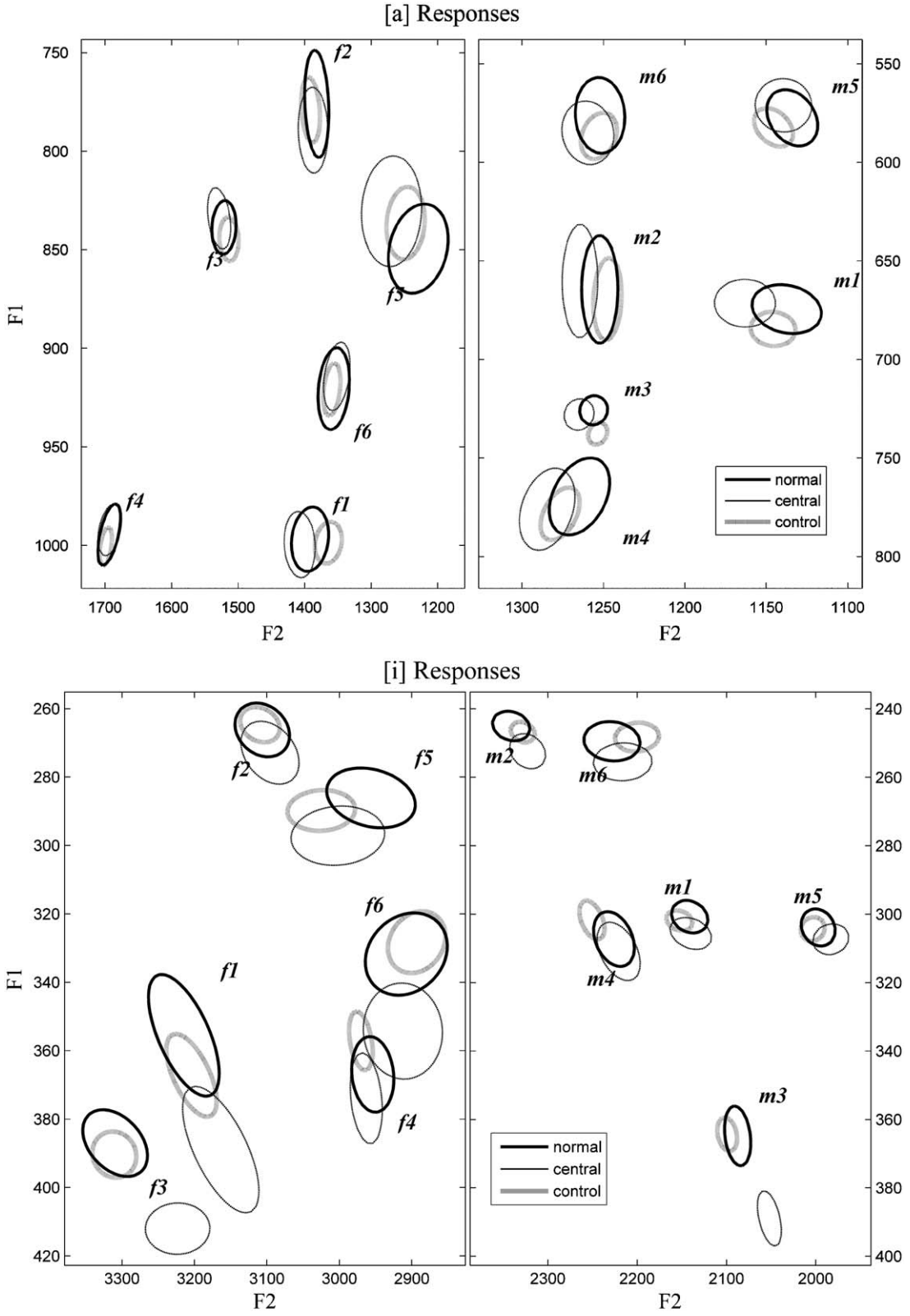

Fig. 4. Confidence regions for [a] responses (top) and [i] responses (bottom) after normal-cue, centralized-cue, and control trials. Ellipses show 95% confidence regions for bivariate means of normal trials (dark bold), centralized-cue trials (thin), and control trials (light thick).

target stimulus. Hence very recent perceptions, on a timescale of roughly 300–1500 ms, can have an impact on production.

The substantial intersubject variation observed in the relations between control and concordant trial responses suggests that subjects may have adopted differing response

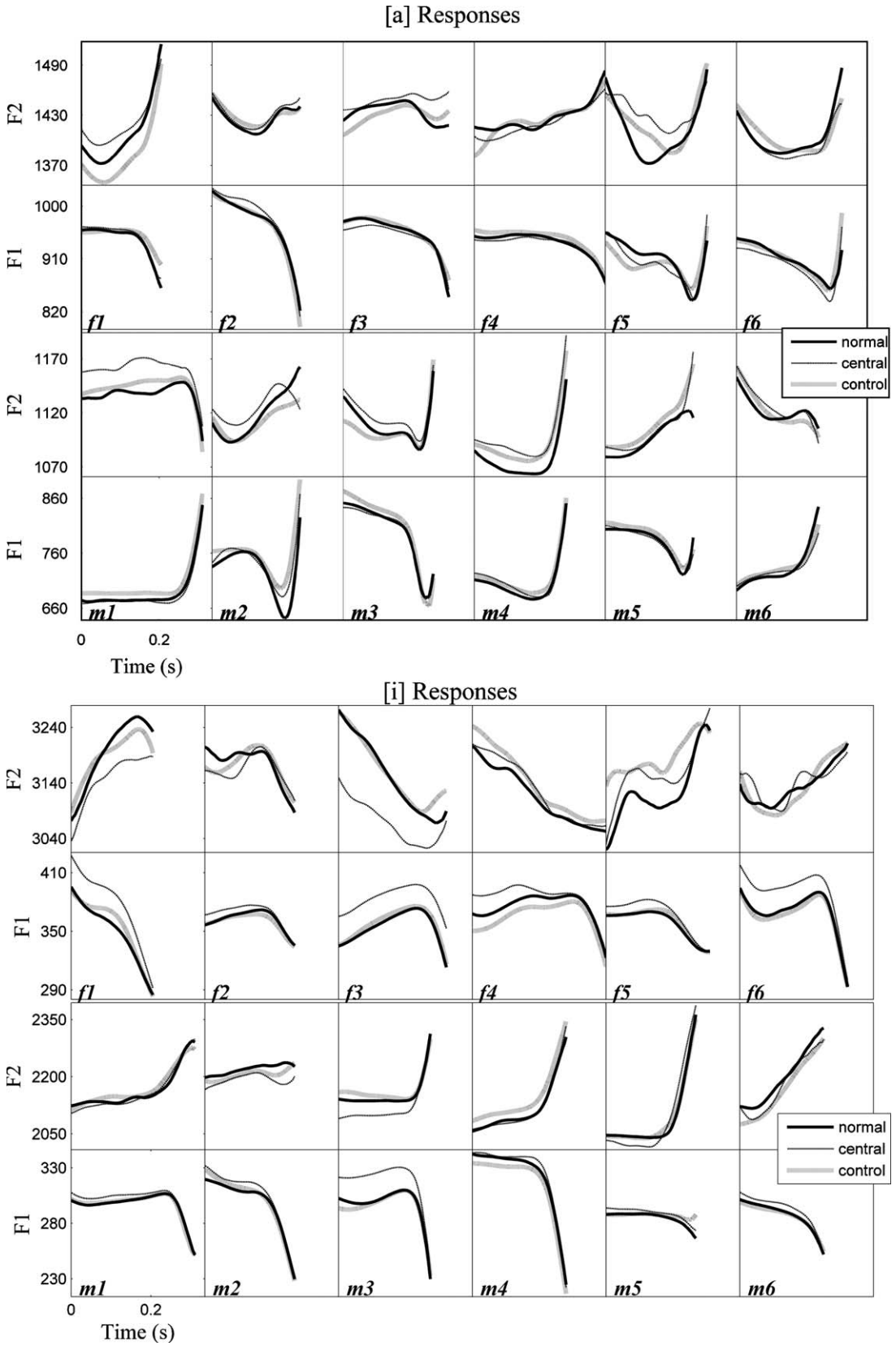

Fig. 5. Average formant trajectories for [a] responses (top) and [i] responses (bottom) after normal (bold), centralized-cue (thin), and control (light thick) trials. All panels in the same row have an identical frequency scale, but different absolute formant ranges.

strategies. It is reasonable to assume that subjects have the ability to pre-plan both vowel responses, prior to and during each trial, and further that subjects control the *extent* of preplanning in order to minimize their response times. It follows that upon hearing the cue vowel, the subject prepares articulatory plans for that same vowel to a greater extent than the non-cue vowel. In contrast, on control trials subjects prepare both vowels to similar extents, because both responses are equally probable. Differences in response preparation may account for fine-grained intersubject variation, but the details of such an account are beyond the scope of the present experimental data.

3.2. Cross-phonemic priming effects

Analyses of variance revealed significant mean F1 and F2 differences between response vowels produced on concordant trials and discordant trials. This suggests that contemporaneous perception and planning of a non-target response influences the planning of the target response. Only results from trials with normal cues are presented here, because the effects of subphonemic detail and cross-phonemic planning are confounded on the discordant centralized-cue trials. The effect of concordance was significant for F2-[a] $\{F(1,1416) = 9.37,\ p < 0.002\}$ and F1-[i] $\{F(1,1430) = 4.02,\ p < 0.045\}$. Interestingly, these were *dissimilatory* effects: vowel formants were less similar to the cue formants on the discordant trials compared with the concordant trials. These effects were even stronger for the mean formants of the first third of each response, where a marginal effect on F2-[i] was observed $\{F(1,1432) = 2.58,\ p < 0.108\}$ in addition to concordance effects on F2-[a] $\{F(1,1418) = 12.07,\ p < 0.001\}$ and F1-[i] $\{F(1,1432) = 11.35,\ p < 0.001\}$.

Fig. 6 shows confidence ellipses for within-subject F1, F2 bivariate means on concordant, discordant, and control trials. Comparing the concordant trials (bold ellipses) to discordant trials (thin ellipses), one can see that a number of subjects exhibited noticeable dissimilatory differences ([a]: *f*1, *f*4, *f*6, *m*1, *m*2, *m*3, *m*4, *m*6; [i]: *f*3, *f*4, *f*5, *f*6, *m*1, *m*3, *m*4). Here, "dissimilation" should be read not in a phonological or historical sense, but in a more literal, phonetic sense, entailing less similarity to the cue vowel: [a] responses tended to be acoustically less like [i] after an [i] cue, and vice versa, [i] responses were less like [a] after an [a] cue.

Where the concordant and discordant trial responses tended to differ, several patterns can be seen in the relations between the control trials and the concordant and discordant trials. One such pattern is for the control trial formants to be relatively more similar to the concordant trials (e.g. [a]: *f*4, *f*6, *m*1, *m*2, *m*4; [i]: *f*2, *f*3, *f*6, *m*1). A comparably common pattern is for the control trial mean vectors to be located either between the concordant and discordant means (e.g. [a]: *f*1, *m*2; [i]: *f*1, *f*5, *m*3), or to be more similar to the discordant trial means (e.g. [a]: *f*5, *m*3, *m*6; [i]: *m*4, *m*6).

Within-subject comparisons of mean F1 and F2 from the first third of responses from concordant and discordant trials showed that most subjects exhibited dissimilatory patterns (cf. the appendix for a table of statistics). For [a]-F1, where there was no main effect of concordance, two subjects showed individually significant differences that were assimilatory in nature; one showed a significant dissimilatory difference. For [a]-F2, 5 subjects had a significant dissimilatory tendency to lower F2 in [a] responses made after an [i], and 10 subjects contributed to the overall trend. For [i]-F1, 4 subjects showed significant dissimilatory patterns, while 9 contributed to the significant trend across subjects. For [i]-F2, 2 subjects showed significant dissimilation.

Fig. 7 shows average F1 and F2 contours from concordant, discordant, and control trials. Although most subjects exhibit some subtle idiosyncrasies, several distinct types of patterns can be seen. First examining the F1 of [a] responses (where higher F1 and lower F2 indicate dissimilation), we see that the majority of subjects had negligible differences between discordant and concordant trials, while two produced on discordant trials somewhat dissimilar responses (*m*1, *m*2) and two others more similar responses (*f*1, *f*5), particularly in the first portion of the vowel. For [a]-F2, the majority pattern was a dissimilatory lowering after discordant cue stimuli (*f*1, *f*4, *f*5, *f*6, *m*1, *m*2, *m*3, *m*4); in some cases this lowering persisted throughout the vowel, in others it occurred primarily in the first portion. In examining [i] responses, lower F1 and higher F2 indicate dissimilation. For [i]-F1, we see that a majority of subjects produced dissimilatory responses on the discordant trials (*f*3, *f*4, *f*5, *f*6, *m*1, *m*3, *m*4), particularly in the first part of the responses. Prime-target dissimilation was observed for these same subjects in [i]-F2 (though less clearly for *f*3). There are two instances in which the discordant trials were associated with noticeable assimilatory changes in F2 (*f*1, *m*6).

One noteworthy aspect of the comparison of formant contours is that there exists some variation with regard to the initial differences between concordant and discordant trial formants as well as the subsequent divergence or convergence of those formants. For example, one pattern is *separation*, where the responses tend to initially differ and remain relatively constantly separated throughout the vowel ([a]-F2: *f*1, *f*4, *f*6; [i]-F2: *f*5, *m*4, *m*6). A more common pattern is *separation-convergence*, where an initial (usually dissimilatory) separation is followed by eventual convergence ([a]-F1: f1, f5; [a]-F2: *m*1, *m*3, *m*4; [i]-F1: *f*1, *f*3, *f*6, *m*3, *m*4; [i]-F2: *m*1, *m*3, *m*4). A third pattern is *proximity-divergence* (*-convergence*), in which formants are initially similar but then eventually diverge ([a]-F1: *m*1, *m*2; [a]-F2: *m*2, *m*6; [i]-F2: *f*2, *m*4, *m*5), and in some instances converge subsequently ([i]-F1: *f*1, *f*5).

Concordance affected response time and duration differently for [a] and [i]. It had an effect on the RT of

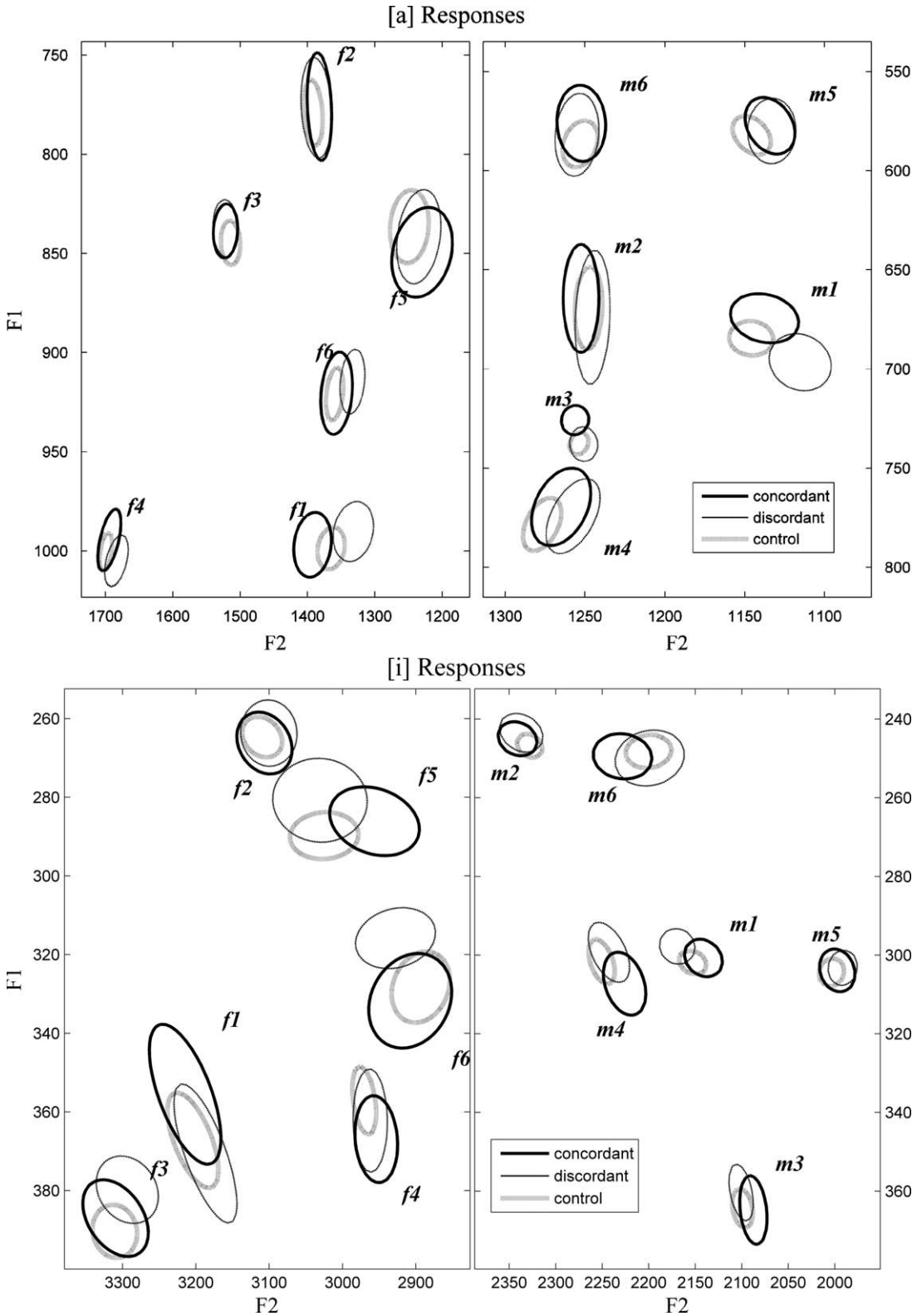

Fig. 6. Confidence regions for [a] response (top) and [i] response (bottom) mean formant vectors from concordant (bold), discordant (thin), and control (thick light) trials. Formants averaged over the first third of each response.

[a] responses $\{F(1,1392) = 6.548,\ p < 0.01\}$, but not [i] responses. [a] responses on concordant trials tended to be made more quickly than on discordant trials. Note that the subjects with the two longest mean response times across the experiment ($f1$, $f5$) were responsible for the only two cases of significant assimilatory differences, which occurred in [a]-F1. Conversely, concordance had a reliable effect on [i] duration $\{F(1,1406) = 6.48,\ p < 0.01\}$, but not on [a]

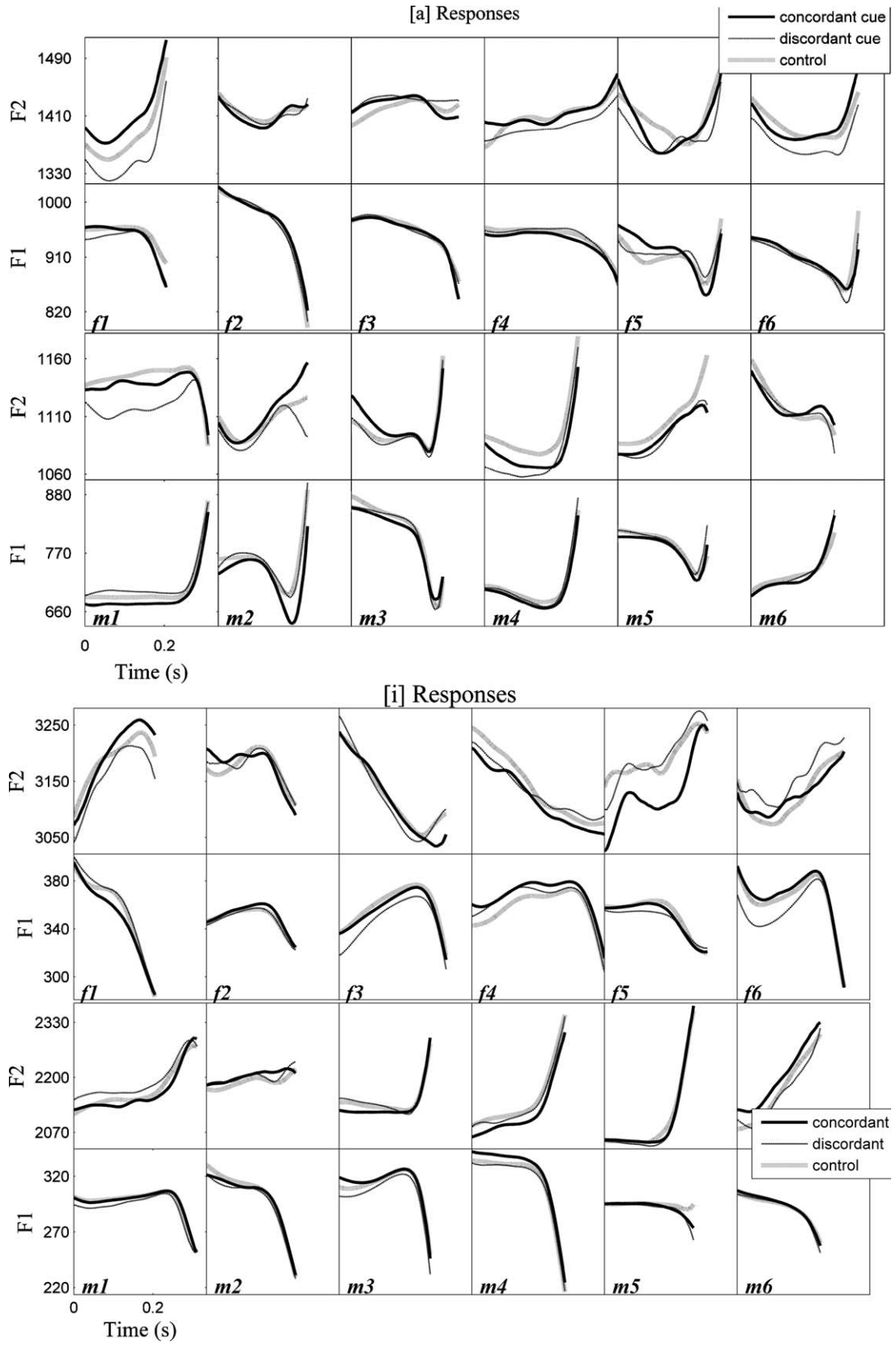

Fig. 7. Average formant trajectories for [a] responses (top) and [i] responses (bottom) after concordant, discordant, and control (normal-cue, no-target) trials. Bold lines show trajectories for concordant trials; thin dashed lines show trajectories for discordant trials; light, thick lines show trajectories for control trials. All panels have the same scale, but different absolute formant ranges.

duration. Examination of within-subject comparisons indicates that the effects on [i]-duration were reliably observed in only two subjects (*m*1, *f*4), who happened to have the two longest mean response durations of all subjects, in the range of 325–400 ms (for comparison, the other subjects exhibited mean response durations in the range of 230–325 ms). Interstimulus delay did not exhibit any significant main or interaction effects on vowel formants between concordant or discordant trials, nor on response times.

Similar dissimilatory patterns were observed on cross-phonemic priming trials with centralized cues, which were not included in the analyses above. For some subjects, dissimilatory discordant trial productions differed between normal and centralized-cue trials. For others, no significant differences were observed. The analysis of these differences is problematic, since they could be attributable either to changes in the mechanism responsible for dissimilation, or to a differential influence of subphonemic detail in cross-phonemic priming. Empirical investigation of how subphonemic memory interacts with cross-phonemic priming is better left for future studies designed to target this interaction.

In sum, a remarkable pattern was observed in the cross-phonemic priming part of this experimental paradigm. For a number of subjects, dissimilatory trends were observed in response formants on discordant trials compared with concordant trials. While this finding begs for an explanation, it would be prudent to conduct further investigation before developing a fully drawn out model. In Section 4.2, a brief sketch of an explanation will be presented, but this sketch does not account for the substantial intersubject variation observed in the cross-phonemic priming. This variation was manifested in several ways: in whether dissimilatory or assimilatory patterns were produced, in the specific vowels and formants that were affected, and in the time course of the effects, which are visible in formant contours.

Notably, there was more variation for some vowel–formant combinations than others. For [a]-F1, significant assimilatory and dissimilatory patterns were seen, albeit in only three subjects. A suggestive correlation here is that the two subjects who produced assimilatory patterns (*f*1, *f*5) were also the two subjects with the longest RTs, and *f*5 also had an anomalously high RT variance—these correlations suggest that the assimilatory patterns produced by *f*1 and *f*5 may have been due to a lack of attention to the task. For [i]-F2, only two subjects produced significant dissimilatory patterns, and the concordance-by-subject interaction was significant. The subject with the largest (but not significant) assimilatory difference in [i]-F2 happened to show an abnormal propensity to respond early, i.e. before the cue. The relations between the assimilatory patterns and RT patterns suggest that either the assimilatory patterns are not general, or that a certain degree of attention is required for the dissimilation to emerge in the task.

The somewhat motley patterns in [a]-F1 and [i]-F2 were matched by robust trends in [a]-F2 and [i]-F1, where all but a few subjects evidenced dissimilation. It thus appears that whatever mechanism is responsible for these dissimilatory patterns, the mechanism can have differing effects on F1 and F2. This could indicate either that F1 or F2 are governed by independent planning mechanisms, or that these variables are not the parameters most directly manipulated by the production system. Note that some subjects exhibited more consistency in dissimilating: for example *m*3 and *m*1 produced significant dissimilatory patterns in 3 and 4 of the vowel–formant combinations, respectively, while others showed in only 1 or 2 cases.

As with the normal vs. centralized-cue comparisons, the control trials did not consistently pattern with either concordant or discordant trials, although the most common pattern was for the controls to be similar to the concordant trials. One problem with the control trials is that by virtue of being interwoven with the experimental trials, subjects may have retained biases that would not otherwise be present in an unbiased two-choice shadowing task. Control trial variation may also have arisen from differences in task strategies adopted by subjects, particularly with respect to balancing maintenance of the cue in working memory with preparation for the two-choice shadowing task. Yet another source of intersubject diversity was variation in the time course of the difference between concordant and discordant mean formant trajectories. Careful observation of these trajectories in Fig. 7 reveals three general patterns of relations: separation, separation-convergence, and proximity-divergence(-convergence). These complex patterns warrant further investigation of cross-phonemic priming.

4. General discussion

The subphonemic priming effects reported in Section 3.1 argue for a speech production model that incorporates episodic memories into production targets. In an exemplar model, the recently perceived the cue vowel can influence a subsequently planned production target. Non-episodic production models, in which targets are abstract categories, do not allow for substantial changes in targets to occur rapidly or be influenced by subphonemic details of recent percepts. Section 4.1 presents two alternative possibilities for when the subphonemic perceptual-motor mapping occurs in the course of a trial, and shows how an exemplar-based model can account for either possibility. Section 4.2 will sketch an explanation for cross-phonemic priming effects, which, although perhaps a more remarkable finding, cannot be unambiguously interpreted as the result of planning processes and episodic memory.

4.1. Subphonemic priming effects in an exemplar model

There are two possible explanations for how subphonemic priming effects arise. The key difference between

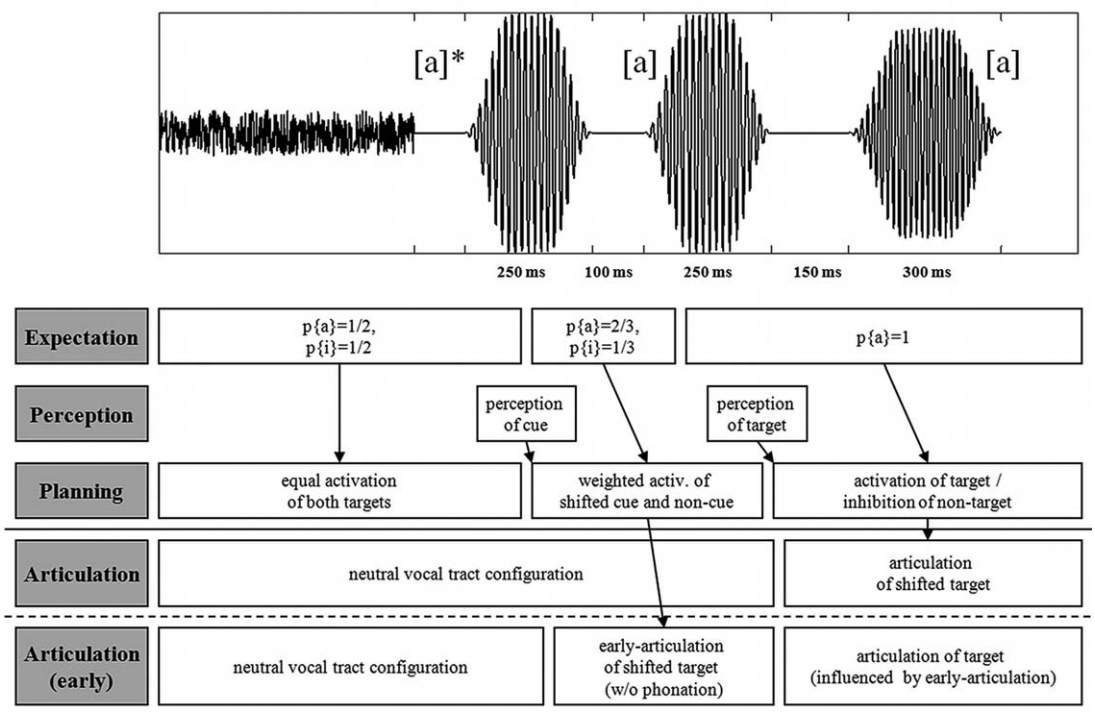

Fig. 8. Schematic representation of the time course of a shifted-cue concordant trial, and hypothesized relations between cognitive processes. (Top) noise, cue [a]*, target [a], and response vowel [a]. Normal and early articulation alternatives are shown.

them is *when* in the course of each trial the mechanisms responsible for the subphonemic priming effects induce articulatory shifts. Fig. 8 presents a simple box-schema of how planning and articulatory processes may occur over the time course of a concordant trial with a centralized cue. Recall that prior to the cue stimulus, the vowels [a] and [i] are equally likely to be the required response. After the cue stimulus, the cue vowel is twice as likely to be the required response. With sufficient exposure to the task, subjects incorporate these probabilities into their pre-trial expectations. Their expectations then change depending upon the cue stimulus, and this affects their response planning. Specifically, after a cue vowel, the activations of the cue and non-cue-vowel targets are weighted proportionally to their associated expectations.

On the one hand, the priming effects may arise from changes exclusively in the planning of speech targets, which are then manifested as articulatory changes immediately prior to (and/or during) the production of the response vowel. The "articulation" level in Fig. 8 represents this interpretation. The activation (or planning) of the response is influenced by the prior weighted activation of a shifted cue vowel. This account holds that before perception of the target stimulus (or perhaps, before the initiation of articulation), the shape of the vocal tract does not vary—in other words, only the speech targets themselves have been altered due to the priming. Experimental effects are then attributed solely to interactions between planning processes, particularly the influence of pre-target planning upon post-target planning. This *planning-interaction* account places the cause of differences between experimental conditions entirely in the realm of speech planning/target selection.

On the other hand, priming effects may arise from changes in vocal tract configuration that take place soon after the cue stimulus has been perceived. Because knowledge of the cue stimulus creates a probabilistic response bias, subjects may pre-shape their vocal tracts, either for the sake of minimizing their response times, or for some other reason (the "early articulation" level in Fig. 8 represents this idea). The pre-configuration of the vocal tract could reflect subphonemic details of the cue stimulus, and would then influence the subsequent production. This *early articulation* account locates the direct cause of the priming effect earlier in the trial.

It is crucial to understand that regardless of which account is correct, subphonemic priming results from the influence of episodic memory upon production. The only difference is precisely when this influence occurs. Because articulatory data were not collected in this experiment, one cannot unambiguously distinguish between the planning interaction or early articulation accounts (in future work articulatory data would be highly informative). Furthermore, these two explanations are not mutually exclusive—both may play a role in generating the priming effects.

Another potential interpretation of the subphonemic priming results is that they might arise from some kind of mimicry effect, rather than episodic memory per se. It is likely that a mimicry effect of this sort involves some of the same memory systems utilized by speech perception and production, in which case this view is hard to distinguish substantively from the one above. It also bears mention in

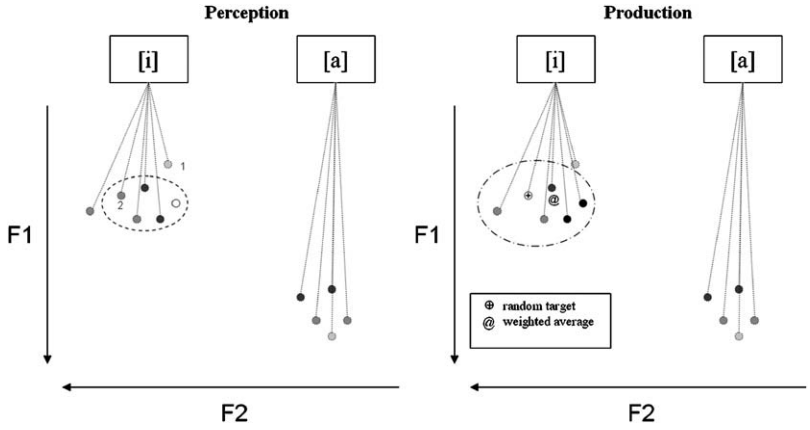

Fig. 9. Simplified schematization of exemplar models of production and perception as described by Johnson (1997) and Pierrehumbert (2001, 2002). (Left) perception model; unfilled circle: new percept; (2) is further from the new percept than (1), but plays a role in perception due to memory-weighting. (Right) production model; + randomly selected target; @ weighted average after entrenchment.

this regard that subjects were not instructed to "imitate" or "mimic" the stimuli, but instead, were instructed to say the correct vowel sound, /a/ or /i/. Hence, if subjects exhibited additional mimicry-like behaviors, these behaviors were not explicitly task-driven.

Fig. 9 presents schematic illustrations of the exemplar models of perception and production described in Pierrehumbert (2001, 2002), which are built upon the exemplar model of perception described by Johnson (1997). The left-hand side of the figure depicts how a new percept (unfilled circle) is categorized according to its similarity to (phonetic distance from) nearby exemplars in a phonetic space. Any number of phonetic dimensions can be used to define a space with a distance metric, but for current purposes we consider only F1 and F2. In categorizing a new percept, phonological category labels (boxes) are activated, and the strength of this activation is the sum of all the activations of the exemplars nearby the new percept. Furthermore, each exemplar has its own base level of activation (represented by the shades of the circles), which decays over time. The effect of this decay is that recent exemplars contribute disproportionately to the categorization of new stimuli. For example, although exemplar_1 is closer to the new percept than exemplar_2, exemplar_1 corresponds to an older memory and in this case fails to contribute to categorization of the stimulus. The most highly activated category then becomes associated with the new percept, which is stored as a new exemplar in memory.

The exemplar-based approach to production is based upon the perception model, but operates in reverse (although cf. Pierrehumbert (2002) for an elaboration of lexical and phonological category influences). The first step in obtaining a production target is for the desired phonological category to be chosen. This "choosing" corresponds to an intention to plan a production target. Next an exemplar associated with the category is randomly selected (Fig. 9, right: +). The probability of an exemplar being chosen is a function of its recency, which reflects the temporal decay of exemplar memories. Then in a process of activation-weighted averaging, the phonetic characteristics of a nearby group of exemplars are averaged. In the Pierrehumbert (2001) model, this group consists of the *n*-nearest exemplars to the randomly selected exemplar (where distance is activation-weighted). The memory-weighting of the distance function makes more recent exemplars more likely than equally distant older exemplars to be integrated into the production target. Additionally, the averaging of phonetic characteristics within the neighborhood of the selected exemplar is memory-weighted, so that more recent exemplars contribute more than older ones to the determination of target values (this is analogous to the role recent exemplars play in the perceptual model).

Now consider how such an exemplar model relates to the box-schema of the time course of perception, planning, and production processes depicted in Fig. 8. For simplicity, we can assume that target planning occurs in the same acoustic exemplar space as perception, which is consistent with the models outlined in Pierrehumbert (2001, 2002). Further assume that talker normalization occurs in the mapping from acoustic to motoric coordinates. (Some plausible alternatives to these assumptions are that planning occurs in a normalized formant space, in motor or gestural coordinates, and/or in some sort of multimodal or amodal coordinates. These complications would not substantially change the basic conceptual structure of the account presented here, but certainly warrant future empirical investigation.)

The exemplar production model operates in the primed-shadowing task as follows. First, prior to the cue, both categories are activated to equal extents and both targets are planned. After the cue has been perceived and stored as an exemplar, two things occur: (1) the activations of the category labels are reweighted according to their likelihood of being the required response and (2) the target is replanned—the fresh exemplar potentially plays a role in determining the phonetic parameters of the new target. Then, after the target stimulus is perceived, the

corresponding vowel becomes fully activated, and a new production target is planned and executed. Articulation and target planning can potentially operate in parallel, meaning that target planning may affect response articulation in mid-production (this provides one way to account for variation in the dynamics of the effects). In other words, even after articulation and phonation have begun, target planning effects may still emerge and disappear.

The planning-interaction account holds that the differences between normal and centralized-cue trials are due to the relatively high activation of the cue exemplar during the planning of the response. Because of its recency, this exemplar contributes disproportionately to target planning. When the cue is an [a]*, the activation-weighted average of [a] exemplars will be biased toward more central formant values, resulting in a more central response.

The early articulation account differs from the planning-interaction account with respect to when and how articulation occurs. This approach holds that following perception of a centralized [a]* cue, the vocal tract is rapidly pre-configured so as to favor the articulation of the [a]*. Presumably, a slight difference between the configurations made after [a] and [a]* cues arises in early articulation due to incorporation of subphonemic details of the cue exemplar. Although normal and centralized cues differed by about 50 and 70 Hz in F1 and F2, activation-weighted averaging of exemplars should tend to decrease the difference, and indeed this is what was observed.

To address whether the early articulation or planning-interaction account should be preferred over the other, consider whether either one makes any predictions that the other does not. The early articulation account relies on the assertion that there are differences in vocal tract geometry between concordant normal and centralized-cue trials (prior to the target stimulus), and that these are maintained up through at least part of the response. Even if we assume a fair degree of nonlinearity in the mapping from acoustics to vocal tract configuration, this account predicts that the formant differences between centralized and normal-cue trials will either decrease or stay constant over the production, depending upon what assumptions are made about the dynamics of target planning. In no case should formant differences increase or exhibit more complex temporal patterns. In other words, the early articulation account predicts that formants will differ most between the two conditions at the beginnings of response contours, or at least will not become substantially more different later in the response.

For the majority of subjects, this prediction of the early articulation account is confirmed, but there are a number of cases where it is violated, too many not to call into question the role of early articulation as the only explanation for the observed priming effects. In these cases ([a]-F1: *f5*, [a]-F2: *f3*, *f5*; [i]-F1: *m2*, [i]-F2: *f1*, *m6*), the normal- and centralized-cue trial formants begin in close proximity, but then diverge such that the centralized-cue responses become more central. In some of these cases, they reconverge. The existence of these patterns suggests that—at least for some subject–vowel–formant combinations—early articulation is not the sole cause of the subphonemic priming effect.

In contrast, the planning-interaction account may accommodate such differences in the following way: individual phonetic values of exemplars can be conceptualized as vectors, specifying values in an acoustic dimension over time (both Johnson (1997) and Pierrehumbert (2001) allow for this). Thus exemplar targets are actually trajectories. It is conceivable in this approach that there exist differences in how production targets are determined, with respect to their temporal dimensions. In other words, subjects may attend to subphonemic details of the cue differently in different phases of the cue. Yet this speculation leaves unanswered questions about why the internal dynamics of the productions are so complex and speaker dependent. In the absence of data to resolve these issues, the question is better left to future investigation.

It is worth noting that variation in which specific formants are affected may arise due to a number of factors, such as individual differences in perceptual ability/attention, memory formation and decay-time, vocal tract geometry, and other linguistic or functional aspects of the vowels involved. The formant-specificity of the centralization effects may be affected by nonlinearities in the perception of acoustic differences. Frequency discrimination is more approximately linear across lower frequencies (<1000 Hz) corresponding to [i]-F1 and [a]-F1, and decays at higher frequencies, so that a 70 Hz difference is more psychoacoustically salient around [a]-F2 than the same difference is around [i]-F2. This may explain why the effect on F2 was observed for more subjects in [a] responses. However, a 50 Hz difference is only slightly more salient for [i]-F1 than the same difference is for [a]-F1, so this cannot explain the general absence of a robust effect on [a]-F1. It is likely that for linguistic-functional reasons, certain formants in speech can bear higher functional loads than others and this influences perceived differences; one could speculate that [a]-F2 is linguistically more salient than [a]-F1 and that this is responsible for the presence/absence of effects in the F2/F1 of [a]—such an explanation would however be pure speculation at this point. Relatedly, higher-level linguistic modulation of perception could lead to a factoring out of variation in acoustic/articulatory dimensions which are correlated with the most salient dimension. In other words, subjects may attribute any centralization perceived in one dimension to centralization perceived in the highly correlated, more salient one. It is also entirely possible that the acoustic measures of F1 and F2 are only indirect associates of some other set of acoustic or motor variables that are the primary targets of the production system.

Regardless of what the ultimate source of such variation may be, subphonemic priming clearly demonstrates that rapid incorporation of subphonemic detail into target planning is possible. Moreover, such phonetic details are

well-understood as aspects of episodic memory, and so we can conclude that episodic memories play a significant role in speech planning and production.

4.2. Cross-phonemic priming effects in an exemplar model

A brief sketch is presented here of one mechanism through which dissimilatory patterns may arise in cross-phonemic priming. However, there are several issues (mentioned below) with this interpretation, and these issues cannot be resolved by the experimental data. For that reason, this sketch is only meant to provide a starting point for future work.

It has been observed in studies of ocular and manual motor behavior that eye movement (saccade) and reaching trajectories to a target location tend to deviate slightly away from non-target distractor locations to which movements have been previously planned (Doyle & Walker, 2001; Sheliga, Riggio, & Rizzolatti, 1994; Van der Stigchel & Theeuwes, 2005). These findings have been modeled by some researchers as arising from the neural inhibition of competing motor plans associated with the non-target saccade or reach (Houghton & Tipper, 1996; Tipper, Howard, & Houghton, 2000). The basic idea is that in order to saccade to or reach for a target, movement plans to other locations must be selectively inhibited. Motor plans to distinct targets are coded by overlapping populations of neurons, and *selective inhibition* of an alternative plan has a small but observable effect upon the executed target movement. Further, the more highly activated the alternative plan(s), the stronger the selective inhibition needs to be.

This model applies fairly straightforwardly to speech target planning in the primed vowel-shadowing task. On a discordant trial, the non-target vowel will be more highly activated than it will be on the concordant trials (because it was the cue stimulus), and hence production of the target vowel will require greater *intergestural inhibition* of the cue. This greater inhibition will shift the distribution of neurons coding for the target vowel away from the cue, resulting in dissimilation relative to concordant trials. In the context of an exemplar model, this can be conceptualized as diminished activation of exemplars most similar to the cue, resulting in a production target that is dissimilated from the discordant cue vowel. Such an effect would not occur between vowels that differ only subphonemically.

A related approach to modeling both the dissimilatory and subphonemic priming effects is to posit that gestural targets are stored in memory as density distributions or distributions of activity in a neural field (e.g. Erlhagen & Schöner, 2002; Guenther, Nieto-Castanon, Ghosh, & Tourville, 2004). In these approaches, episodic memories alter the activation dynamics of the neural field, both on working memory and long-term memory timescales, and interesting effects can emerge when multiple movements are contemporaneously planned. For example, Erlhagen and Schöner (2002) present a dynamical field model in which nearby reaching targets are integrated in target planning. This result is analogous to subphonemic priming: the cue evokes an activation pattern very similar to, but subtly different from, the activation pattern evoked by the target—the result from integrating these activation patterns is a movement target that is a compromise between them. Houghton and Tipper (1996) and Tipper et al. (1999) hypothesized that when the contemporaneously planned targets are distinct enough, inhibition is used to select the appropriate target. This assumption is consistent with results from Ghez et al. (1997), who showed that for reaching movements, the proximity of the non-target movement to the target movement determines whether deviation of the reach trajectory towards or away from the non-target is observed. Dynamical field approaches are conceptually distinct from exemplar-based models, but share the important property that perceptions can substantially alter the memory that is used to determine production targets.

An alternative interpretation of the dissimilation pattern is that articulator movement amplitude is diminished on the concordant trials due to early articulation of the cue. If the vocal tract is biased toward production of the cue prior to the target stimulus, movement amplitude and velocity may be diminished, the effect being hypoarticulation on the concordant trials. Likewise on discordant trials, if the vocal tract is pre-shaped with a cue-vowel bias, movement amplitude and velocity may increase, resulting in hyper-articulation. One or both of these effects could explain the dissimilatory patterns, and do so in a way that would involve no difference in target planning between the concordant and discordant trials.

The most reliable way to distinguish the inhibitory and early articulation accounts would be to collect articulatory data during the task, but since this was not done in the present experiment, no definitive claims should be made. The issue is worth resolving in future work, because intergestural inhibition, if it exists, may inform our understanding of a variety of linguistic patterns, including boundary effects on gestural amplitude and duration, neighborhood density effects, and diachronic forces such as contrast preservation.

Potentially related to the dissimilatory patterns in the present study is the phenomenon of compensation for altered auditory feedback (Houde & Jordan, 2002; Larson, Altman, Liu, & Hain, 2008; Purcell & Munhall, 2006). Generally speaking, in altered auditory feedback experiments, speakers hear a version of their own speech that has been manipulated in some way. For example, Houde and Jordan (2002) shifted vowel formants so that speakers heard either /a/ or /i/ when they produced /ɛ/. Speakers compensated to varying degrees for this altered feedback by shifting the targets of their vowel productions. Furthermore, when feedback was subsequently blocked by noise, the compensatory shifts were partly retained. These findings indicate that perceptual feedback is integrated into memory and subsequently influences

Table A.1
Within-subject F1 and F2 comparisons between normal- and centralized-cue trials.

Subject	F1						Subject	F2					
	Normal		Central		Normal-centralized			Normal		Central		Normal-centralized	
	Hz	(σ, N)	Hz	(σ, N)	Δ	$p<$		Hz	(σ, N)	Hz	(σ, N)	Δ	$p<$
/a/							/a/						
*f*5	849	(62,49)	830	(77,49)	−19	0.09^{+}	*f*6	1356	(64,47)	1352	(55,50)	−5	0.65
*m*5	577	(42,54)	571	(39,52)	−6	0.21	*f*4	1694	(47,48)	1693	(49,50)	−1	0.53
*f*6	921	(56,47)	914	(48,50)	−6	0.28	*m*5	1134	(45,54)	1139	(49,52)	5	0.28
*f*3	839	(37,48)	834	(43,50)	−5	0.29	*f*2	1382	(64,78)	1387	(78,77)	6	0.32
*m*2	664	(96,77)	660	(100,76)	−4	0.40	*f*3	1521	(50,48)	1529	(49,50)	7	0.23
*m*1	674	(43,76)	671	(43,77)	−3	0.33	*m*6	1252	(41,48)	1260	(44,49)	8	0.20
*f*4	994	(43,48)	992	(36,50)	−2	0.39	*m*3	1256	(31,80)	1265	(33,80)	9	0.04*
*m*3	726	(27,80)	728	(29,80)	2	0.69	*m*2	1252	(39,77)	1264	(38,76)	12	0.03*
*f*1	957	(54,68)	960	(56,69)	3	0.61	*f*1	1392	(92,68)	1407	(77,69)	15	0.15
*m*4	770	(55,51)	776	(57,49)	6	0.72	*m*4	1265	(52,51)	1285	(47,49)	20	0.02*
*m*6	576	(52,48)	585	(44,49)	9	0.81	*m*1	1137	(74,76)	1163	(66,77)	26	0.01*
*f*2	776	(96,78)	789	(76,77)	13	0.83	*f*5	1229	(124,49)	1269	(124,49)	40	0.06^{+}
/i/							/i/						
*m*5	304	(16,53)	307	(13,51)	3	0.12	*f*3	2810	(123,49)	2723	(122,49)	−86	0.01*
*m*4	307	(23,52)	311	(24,52)	4	0.22	*f*1	3215	(165,72)	3164	(177,71)	−51	0.04*
*m*1	301	(17,81)	306	(17,82)	5	0.04*	*m*3	2088	(52,78)	2052	(48,79)	−36	0.01*
*m*6	249	(16,48)	255	(15,48)	6	0.03*	*m*2	2341	(73,76)	2323	(70,76)	−18	0.06^{+}
*m*2	245	(15,76)	252	(18,76)	7	0.01*	*m*5	1997	(55,53)	1983	(57,51)	−14	0.10^{+}
*f*2	286	(28,79)	293	(33,80)	7	0.09^{+}	*m*6	2228	(86,48)	2216	(90,48)	−12	0.25
*f*4	367	(30,49)	374	(37,50)	7	0.15	*f*2	3106	(134,79)	3096	(143,80)	−11	0.32
*f*5	286	(23,46)	297	(24,49)	11	0.01*	*m*4	2126	(65,52)	2121	(69,52)	−5	0.34
*f*6	332	(33,50)	354	(39,50)	23	0.01*	*m*1	2141	(74,81)	2140	(85,82)	−1	0.46
*m*3	365	(31,78)	389	(29,79)	24	0.01*	*f*6	3008	(157,50)	3012	(152,50)	5	0.56
*f*3	387	(27,49)	412	(21,49)	25	0.01*	*f*4	2954	(80,49)	2963	(62,50)	9	0.73
*f*1	356	(60,72)	389	(62,71)	33	0.01*	*f*5	2956	(162,46)	3002	(177,49)	46	0.90

One-sided *t*-tests.
* >95% confidence in a significant difference between population means.
$^{+}$ >90% confidence.

production targets. It is possible—although admittedly speculative—to see such compensatory responses as a form of intergestural inhibition. In that case, the memory of the perception of altered feedback could have an effect similar to the planning of a non-target response; the non-target exemplars associated with the altered auditory feedback would be more strongly inhibited than usual, resulting in dissimilation of the speech target away from phonetic values associated with the altered feedback.

5. Conclusion and future directions

The results observed in this primed vowel-shadowing experiment showed clear effects of subphonemic priming on vowel formants. These effects argue that episodic memory plays a role in the formation of speech targets, and they are well-understood in the context of exemplar models of perception and production (Johnson, 1997; Pierrehumbert, 2001, 2002). Subphonemic details of a cue stimulus are stored in memory as an exemplar, i.e. a set of associations between phonetic values and various linguistic and non-linguistic categories. The recency of the cue-vowel exemplar endows it with a relatively large influence over the subsequently planned production target, resulting in relatively centralized production after a centralized-cue vowel. Furthermore, this priming effect can happen rapidly, on timescales as short as approximately 300 ms.

Cross-phonemic priming effects were primarily dissimilatory, with discordant trial response vowel qualities being less similar to the prime vowel than response vowel qualities on concordant trials. While some form of intergestural inhibition may be useful in understanding these and other speech patterns, an account based upon early articulation cannot be ruled out. Only further experimental work that measures articulation directly can address this issue.

Primed shadowing is a method for investigating cognitive planning mechanisms involved in speech, and in the present case has demonstrated that effects of recent percepts can exert substantial influences on subphonemic details of articulation. These findings indicate that speech production involves episodic memory, and confirm predictions of exemplar production models. This experimental paradigm has also provided some intriguing cross-phonemic priming results that merit further exploration.

Table A.2
Within-subject formant comparisons between concordant and discordant trials.

Subject	F1						Subject	F2					
	Concordant		Discordant		Discordant–concordant			Concordant		Discordant		Discordant–concordant	
	Hz	(σ, N)	Hz	(σ, N)	Δ	$p<$		Hz	(σ, N)	Hz	(σ, N)	Δ	$p<$
[a]							[a]						
*f*5	876	(64,49)	850	(72,47)	−26	0.07^{+}	*f*1	1375	(101,69)	1332	(86,69)	−43	0.01*
*f*1	960	(51,69)	943	(49,69)	−18	0.04*	*f*5	1284	(119,49)	1258	(106,47)	−26	0.26
*f*6	952	(58,48)	945	(59,47)	−7	0.58	*m*1	1133	(73,76)	1113	(62,79)	−20	0.07^{+}
*m*6	567	(50,47)	563	(55,44)	−5	0.67	*f*4	1690	(57,49)	1672	(66,49)	−17	0.17
*f*2	803	(83,78)	806	(87,79)	2	0.86	*m*4	1277	(50,51)	1261	(47,50)	−16	0.10^{+}
*m*3	749	(29,82)	753	(36,79)	4	0.42	*m*3	1275	(38,82)	1261	(38,79)	−15	0.02*
*f*3	851	(34,50)	856	(41,48)	5	0.51	*f*6	1376	(71,48)	1363	(65,47)	−14	0.34
*m*4	793	(62,51)	798	(58,50)	5	0.69	*m*2	1239	(42,77)	1233	(35,73)	−6	0.35
*f*4	994	(38,49)	1001	(32,49)	7	0.30	*m*5	1121	(39,53)	1119	(39,49)	−1	0.87
*m*5	588	(42,53)	596	(50,49)	8	0.39	*m*6	1272	(47,47)	1273	(36,44)	1	0.88
*m*2	668	(97,77)	683	(107,73)	15	0.38	*f*3	1512	(54,50)	1515	(54,48)	3	0.81
*m*1	674	(47,76)	695	(50,79)	21	0.01*	*f*2	1403	(67,78)	1410	(81,79)	7	0.56
[i]							[i]						
*f*6	331	(34,50)	308	(22,49)	−23	0.00*	*m*6	2163	(102,48)	2136	(101,48)	−27	0.19
*f*3	364	(33,50)	352	(28,49)	−13	0.04*	*f*1	3118	(151,72)	3097	(147,69)	−21	0.40
*m*3	361	(32,77)	348	(31,80)	−13	0.01*	*f*2	3100	(189,79)	3089	(179,76)	−11	0.70
*m*4	312	(20,52)	303	(19,51)	−9	0.02*	*m*5	1999	(60,54)	1996	(49,53)	−3	0.79
*m*1	298	(18,81)	292	(18,78)	−5	0.07^{+}	*m*2	2322	(54,76)	2320	(77,73)	−2	0.82
*f*4	353	(29,49)	350	(33,48)	−3	0.66	*f*4	3025	(63,49)	3034	(53,48)	9	0.45
*f*5	286	(30,46)	282	(29,48)	−3	0.57	*f*3	2904	(132,50)	2916	(122,49)	12	0.64
*m*2	253	(20,76)	251	(19,73)	−2	0.45	*m*4	2103	(74,52)	2125	(62,51)	22	0.11
*f*2	277	(24,79)	275	(21,76)	−2	0.60	*m*1	2129	(77,81)	2157	(76,78)	28	0.02*
*m*6	257	(15,48)	256	(22,48)	−1	0.77	*m*3	2088	(56,77)	2117	(64,80)	29	0.01*
*m*5	305	(19,54)	305	(16,53)	0	0.95	*f*6	2994	(179,50)	3023	(165,49)	29	0.41
*f*1	378	(60,72)	389	(59,69)	10	0.30	*f*5	2931	(185,46)	2991	(189,48)	60	0.13

Two-sided *t*-tests.
*95% confidence in a significant difference between population means.
$^{+}$90% confidence.

Acknowledgements

Many thanks to Keith Johnson, Rich Ivry, and Sharon Inkelas for advice and discussion at every stage of this research. I also wish to thank Louis Goldstein and two anonymous reviewers for numerous helpful comments in the authoring of this manuscript. Thanks to Ron Sprouse, Christian Dicanio, Grant McGuire, and Molly Babel for various assistance in the UC Berkeley Phonology Lab. Special thanks to Kim and Cade Tilsen.

Appendix

Table A.1 shows within-subject comparisons of normal and centralized-cue trial formants conducted using 1-sided *t*-tests. Table A.2 shows within-subject comparisons of the mean F1 and F2 of responses from concordant and discordant trials, where formant measurements were taken from the first third of each response. Confer Sections 3.1 and 3.2 for discussion of these patterns.

References

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Doyle, M., & Walker, R. (2001). Curved saccade trajectories: Voluntary and reflexive saccades curve away from irrelevant distractors. *Experimental Brain Research*, *139*, 333–344.

Erlhagen, W., & Schöner, G. (2002). Dynamic field theory of movement preparation. *Psychological Review*, *109*(3), 545–572.

Fowler, C., Brown, J., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language*, *43*(3), 396–413.

Ganong, W. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance*, *6*, 110–125.

Ghez, C., Favilla, M., Ghilardi, M. F., Gordon, J., Bermejo, R., & Pullman, S. (1997). Discrete and continuous planning of hand movements and isometric force trajectories. *Experimental Brain Research*, *115*, 217–233.

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

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

Goldinger, S., Pisoni, D., & Logan, J. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(1), 152–162.

Guenther, F., Nieto-Castanon, A., Ghosh, S., & Tourville, J. (2004). Representation of sound categories in auditory cortical maps. *Journal of Speech, Language, and Hearing Research*, *47*, 46–57.

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

Hintzman, D., Block, R., & Inskeep, N. (1972). Memory for mode of input. *Journal of Verbal Learning and Verbal Behavior*, *12*, 741–749.

Houde, J., & Jordan, M. (2002). Sensorimotor adaptation of speech I: Compensation and adaptation. *Journal of Speech, Language, and Hearing Research*, *45*, 295–310.

Houghton, G., & Tipper, S. (1996). Inhibitory mechanisms of neural and cognitive control: Applications to selective attention and sequential action. *Brain and Cognition*, *30*, 20–43.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson, & J. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego: Academic Press.

Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, *34*, 485–499.

Larson, C., Altman, K., Liu, H., & Hain, T. (2008). Interactions between auditory and somatosensory feedback for voice F0 control. *Experimental Brain Research*, *187*, 613–621.

Liberman, A., & Mattingly, I. (1985). The motor theory of speech perception revised. *Cognition*, *21*, 1–33.

Oudeyer, P.-Y. (2006). *Self-organization in the evolution of speech: Studies in the evolution of language*. Oxford: Oxford University Press.

Palmeri, T., Goldinger, S., & Pisoni, D. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(22), 309–328.

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

Pierrehumbert, J. (2002). Word-specific phonetics. *Laboratory Phonology*, *VII*, 101–139 Mouton de Gruyter, Berlin.

Pierrehumbert, J. B. (2004). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, 115–154.

Pierrehumbert, J. (2006). The next toolkit. *Journal of Phonetics*, *34*, 516–530.

Purcell, D., & Munhall, K. (2006). Compensation following real-time manipulation of formants in isolated vowels. *Journal of the Acoustical Society of America*, *119*(4), 2288–2297.

Raganath, C., Johnson, M., & D'Esposito, M. (2003). Prefrontal activity associated with working memory and episodic long-term memory. *Neuropsychologia*, *41*, 3778–3789.

Sheliga, B., Riggio, L., & Rizzolatti, G. (1994). Orienting of attention and eye movements. *Experimental Brain Research*, *98*, 507–522.

Suprenant, A. M., & Neath, I. (2008). The nine lives of short-term memory. In A. Thorn, & M. Page (Eds.), *Interactions between short-term and long-term memory in the verbal domain* (pp. 16–33). London: Taylor & Francis.

Tipper, S., Howard, L., & Houghton, G. (2000). Behavioral consequences of selection from neural population codes. In S. Monsell, & J. Driver (Eds.), *Control of cognitive processes* (pp. 225–245). Cambridge, MA: MIT Press.

Van der Stigchel, S., & Theeuwes, J. (2005). The influence of attending to multiple locations on eye movements. *Vision Research*.

Wedel, A. (2004). Category competition drives contrast maintenance within an exemplar-based production/perception loop. In *Proceedings of the seventh meeting of the ACL special interest group in computational phonology* (pp. 1–10). Barcelona, Spain: Association for Computational Linguistics, July 2004.

Yaegor-Dror, M. (1996). Phonetic evidence for the evolution of lexical classes: The case of a Montreal French vowel shift. In G. Guy, C. Feagin, J. Baugh, & D. Schiffrin (Eds.), *Towards a social science of language* (pp. 263–287). Philadelphia: Benjamins.

Yaegor-Dror, M., & Kemp, W. (1992). Lexical classes in Montreal French. *Language and Speech*, *35*, 251–293.