TR - H - 154

Experiments in Vowel Segregation

Alain de Cheveigné

1995. 7. 11

ATR人間情報通信研究所

〒619-02　京都府相楽郡精華町光台2-2　☎ 0774-95-1011

ATR Human Information Processing Research Laboratories

2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

Telephone: +81-774-95-1011

Facsimile: +81-774-95-1008

EXPERIMENTS

IN VOWEL SEGREGATION.

Alain de Cheveigné

1. Abstract

This report describes a series of experiments on the segregation of mixed vowels, using techniques designed to improve sensitivity to segregation cues. The first experiment was to test and calibrate these techniques. It provided detailed information on how identification depends on combined factors of relative level and F0 difference. We found that an inter-vowel level mismatch, combined with a task in which the subject is free to answer *one or two* vowels, can greatly enhance the sensitivity of the double-vowel identification paradigm.

The second experiment investigated the possibility that the $\Delta F0$ effects observed in classic "double vowel" studies might be conditional on the phase patterns of the vowels employed, as recent theories of vowel segregation based on temporal beat patterns and pitch period asynchrony might lead us to believe. We found little evidence that intra-vowel or inter-vowel phase patterns determine segregation, whether at unison or at a $\Delta F0$ of 6%.

The third experiment investigated more precisely whether phase effects could have caused an artifact in a previous experiment on harmonicity. We found no evidence of such an artifact.

The fourth experiment reinforced this conclusion by replicating three main conditions of the harmonicity experiment with stimuli designed to minimize eventual phase or beating effects. We found, as previously, a strong dependency of identification on the harmonicity of the ground (interfering) vowel, consistent with the hypothesis of harmonic cancellation. However we no longer observed the paradoxical effect of target harmonicity (opposite to that predicted by the harmonic enhancement hypothesis) that we had found previously. Target harmonicity had no measurable effect.

The fifth experiment replicated several conditions of the previous experiments using a more classic task. Our one-or-two response task is sensitive to cues that signal the *multiplicity* of sources within a stimulus, whereas the classic two vowel forced response task ignores these cues, and is mainly sensitive to cues that determine *mutual masking* between vowels. Replication with the same subjects and conditions allowed us to assess the impact of the new task, and to establish a basis of comparison with prior results. As expected, we found smaller effects with the classic task, but overall patterns were similar.

The sixth experiment was a full replication our previous harmonicity experiment, using the classic two-response task. We found as before a strong effect of background harmonicity, but no effect of target harmonicity. The results once more support the cancellation hypothesis but not the enhancement hypothesis.

The seventh experiment investigated two conditions in which the target and background were both harmonic and had the same nominal F0. In one condition the inharmonic patterns were identical, in the other they were different. Identification was poor in the first case, as for harmonic stimuli at unison. It was better in the second case. No conclusion of interest is drawn from this result.

2. Introduction

It is well known that when several talkers speak together, differences in fundamental frequency (F_0) between voices make the speech of each talker easier to understand. Fig. 1, modified from de Cheveigné, et al. (1995), summarizes some of the classic results obtained in experiments using pairs of synthetic vowels.

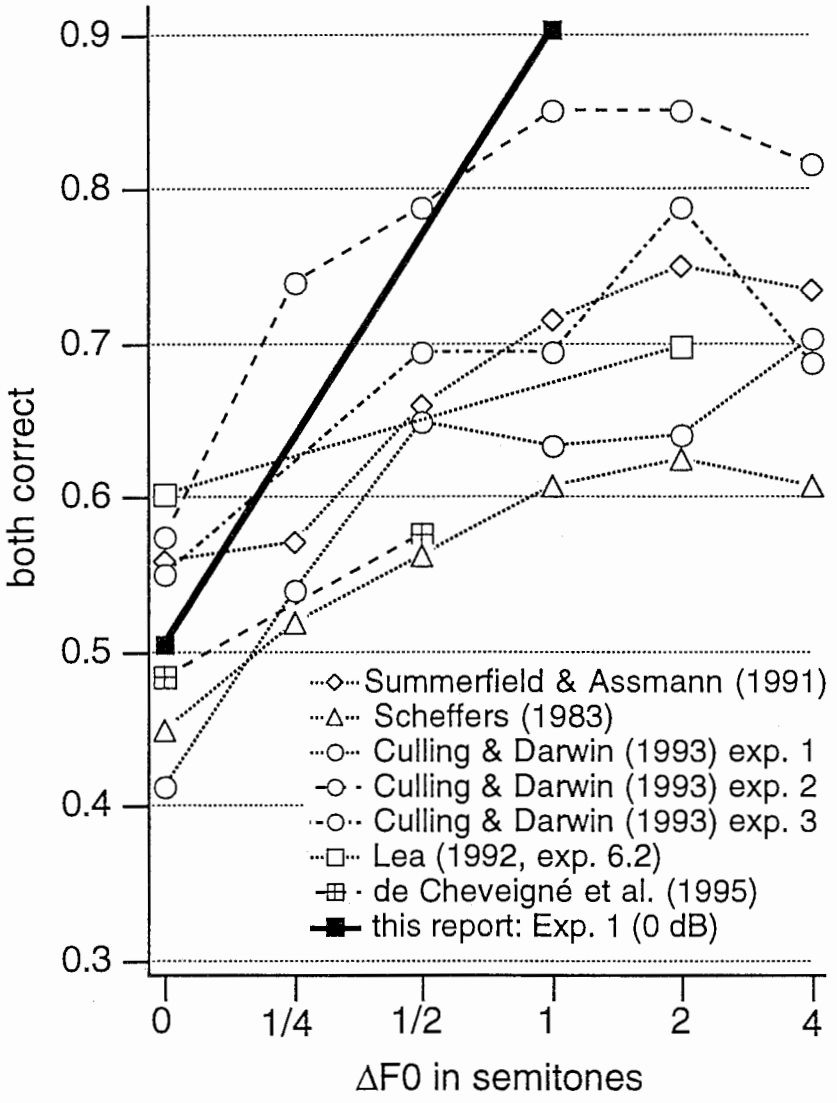


Fig. 1. Combination-correct identification rates as a function of ΔF_0 reported in previous studies (dotted lines), and obtained in the present study (continuous line).

These results all show that identification gets better when the F_0s of the two vowels are made different. Several models and methods have been proposed to explain F_0-guided segregation, or reproduce it within interference-reduction systems (see de Cheveigné, 1993a) for a review. In our recent work we have tried to clarify whether "F_0-guided" segregation depends on the F_0 of the *target* (or more precisely, its harmonicity), or that of the *background*. This alternative corresponds to a choice between two segregation mechanisms that we call harmonic *enhancement* and harmonic *cancellation*, respectively. Accounts of auditory scene analysis often invoke harmonicity as a sort of "glue" that keeps

together partials that belong to the same voice. This is a form of harmonic enhancement. However there is evidence that cancellation is more effective for real speech (de Cheveigné, 1993b, 1994; de Cheveigné et al. 1994a). There also evidence that it is actually employed by the auditory system in situations such as double-vowel experiments (de Cheveigné et al. 1994, 1995; Summerfield and Culling, 1992b; Culling et al., 1994; Lea, 1992). Evidence in favor of enhancement is so far much weaker.

However, recently other mechanisms have been suggested, based on aspects of the temporal patterns that co-vary with ΔF_0. These mechanisms do not directly involve harmonicity as in F_0-guided segregation. As such, they escape our dichotomy of enhancement vs. cancellation. Temporal patterns are phase dependent, and this raises the question of whether the segregation effects observed in double-vowel experiments are conditioned by the particular phase patterns employed.

In planning the following experiments, we were particularly concerned by the possibility that the phase patterns that occurred in stimuli of our experiments on harmonicity (de Cheveigné et al., 1995) might have produced some of the effects that we attributed to harmonicity. Our primary goal was to clarify this question.

3. General methods

3.1. Stimuli , presentation and subjects

Stimuli for all experiments consisted in either single or double synthetic vowels representing the set of Japanese vowels /a/, /i/, /u/, /e/, /o/. Details of synthesis are given in Appendix A. Stimuli were presented via headphones (Stax SR-Λ), at a sound pressure level of 63 to 70 dBA. The sound system was calibrated using a Bruel&Kjaer artificial ear (sound level meter type 2231, half-inch microphone type 4134). Subjects were seated in a sound-treated booth or room, facing a computer terminal that was used to give prompts and gather responses. Subjects were six native speakers of Japanese, two male and four female, aged 18 to 27. Two belonged to ATR staff, and four were students paid for their services.

3.2. Task and scoring

In all experiments (except the last three), subjects were presented once with each stimulus and requested to answer either one or two vowels. They were informed that the stimuli could contain one or two vowels, that the vowels belonged to the set /a/, /e/, /i/, /o/, /u/, and that, in the case of two vowels, the vowels within a pair were different. They were told that the vowels were synthetic and might sound strange, and that in some cases they might not be intelligible. They had the possibility to answer "x" instead of a vowel that they could not identify. They could pause at will, in which case the last stimulus before the pause was presented again after the pause. A session typically lasted between one and two hours.

Single vowel stimuli were scored once: the response was considered correct if the response contained the name of the vowel (regardless of whether the subject responded one or two vowels). Double vowel stimuli were scored twice, once for each vowel. The response for each vowel was classified according to its nature (phoneme, F_0, phase, harmonicity), the nature of the second vowel, and their eventual relationship (ΔF_0, relative level). For all stimuli, the number of vowels responded was noted.

This scoring method provides "constituent correct" scores, and differs from the more familiar method of counting responses in which both vowels are correct ("combination-correct" scores, as in Fig. 1). Our method doubles the number of responses that can be exploited, and allows scores to be calculated separately for each vowel as a function its state (frequency, phase, harmonicity, etc.) and the state of the vowel that is mixed with it. One can thus measure how segregation depends on characteristics of target and background (Lea, 1992; de Cheveigné et al., 1995).

3.3. Design

Each of the six subjects performed five sessions with the stimulus set, on five days. This design was chosen preferably to a design with more subjects but fewer sessions because we expected strong individual differences. Such differences are possibly of interest in themselves, but they may weaken the power of a repeated measures analysis (which tests for the generality of conclusions over the population sampled by the subjects). Given the present design, we can interpret the significance of an effect at either of two levels of generality:

a) the effect is robust over the population (as judged by a repeated measures analysis with random factor subject),

b) the effect is significant for at least one subject (as judged by an overall fixed-effects ANOVA, or a fixed-effect ANOVA for that subject), but possibly too different between subjects to allow inferences at the population level.

The latter sort of interpretation is important if we wish to give weight to the finding that an effect is *not* significant.

The seven experiments described here were carried out in three stages. First stage was Experiment 1, from which was derived the level mismatch applied in following experiments. The second stage comprised Experiments 2-4, and the third stage comprised Experiments 5-7. Stimuli for experiments within a stage were presented together within each session, in interleaved fashion. Subjects performed a total of 15 sessions altogether.

4. Experiment 1

4.1. Introduction

Experiment 1 is a preliminary experiment designed to determine an appropriate inter-vowel level to eliminate ceiling effects and improve the sensitivity of the double-vowel identification paradigm.

A problem often noted in double vowel experiments is the small size of effects. This may be due in part to ceiling effects: identification at unison is perfect for certain subjects and/or vowel pairs, leaving no room for improvement with F0 differences. De Cheveigné et al. (1995) reasoned that this might happen for one vowel within a pair if there was a serious level mismatch, and they tried to determine corrective level factors to balance mutual masking. However that reasoning was flawed: there is no guarantee that, once levels are balanced, *both* vowels won't suffer ceiling effects. Here, we apply on the contrary a systematic level imbalance, to avoid the region in which identification is at a ceiling. Experiment 1 was designed to test the idea and determine appropriate level factors to use in subsequent experiments.

Classic double vowel experiments require subjects to answer two vowels for every stimulus. This has several consequences: a) the task is uncomfortable when only one vowel can be heard, b) the subject may use a particular vowel as a default response, thus unwittingly scoring perfect identification for that vowel, c) segregation cues that signal the *multiplicity* of sources are ignored, d) the subject is under pressure to improve her performance, so there may be training effects. It seems that training may contribute to reduce the size of effects (Assmann and Summerfield, 1994). Instead of requiring two vowel responses, we told our subjects that the stimuli contained either single or double vowels, and we requested them to answer either one or two vowels. The number of vowels responded is in itself an interesting measure.

4.2. Methods

Single vowels were synthesized in Klatt phase at frequencies of 124 Hz and 132.5 Hz (see Appendix A for details). Double vowels consisted of two vowels with the same F0 ($\Delta F0 = 0\%$) or different F0s ($\Delta F0 = 6\%$), scaled to obtain a level offset of -20, -10, 0, 10, or 20 dB, and added. The sum was then scaled to a fixed RMS. level. Stimuli were 200 ms in duration with 20 ms raised cosine onset and offset ramps.

Double vowel conditions within a stimulus set were: (10 unordered vowel pairs) x (5 levels) x (2 $\Delta F0$s) x (2 F0 orders) x (3 repetitions) = 600 double-vowel stimuli. To these were added 240 single vowel stimuli (5 vowels) x (2 F0s) x (24 repetitions). A relatively large proportion of single vowels was included to ensure that the stimulus set was as described to the subjects. It also allowed us to check single vowels for possible effects of synthesis parameters (F0, phase, harmonicity) on vowel quality. Stimulus order was randomized for each session. Sessions generally required between one and two hours to complete, including pauses. Each subject performed five sessions on separate days.

4.3. Results

4.3.1. Single vowels

Identification rates were calculated for every vowel, frequency, subject, and session. Each data point was based on 24 responses. Overall identification rate was 99.75%. The lowest rate for a vowel was 99.2% (/i/) and the lowest rate for a subject was 99.3% (N). Evidently, subjects had no difficulty identifying the synthetic vowels. About 10% of all single vowels evoked two-vowel responses, with considerable differences between subjects (27% for K, 2% for U), but only small differences between vowels, and no effect of frequency.

4.3.2. Double vowels

4.3.2.1. Statistical analysis

The data were analyzed in several steps:

1) Constituent-correct identification rates were calculated for each vowel pair, level mismatch, F_0, ΔF_0, subject and session, averaged over repetition. Each data point represented three responses. A preliminary fixed-effects ANOVA was performed with factors PAIR, F_0, ΔF_0 and SUBJECT, and first order interactions. No effect or interaction involving F_0 was significant. In other words: at unison it made no difference to any subject whether the vowels were at 125 or 132.5Hz; at a ΔF_0 of 6% it made no difference if the target was at 125 Hz and the background at 132.5 Hz, or vice-versa. Given the various ways in which F_0 can interact with formant structure, and given the sensitivity of our methods to other factors, this result is perhaps surprising.

2) Scores were averaged over F_0 and transformed according to the formula arcsin(2*rate-1) to make distributions more homogenous. Relative levels of +10 and +20 dB were eliminated from analysis, as they were strongly affected by ceiling effects. The scores were submitted to a repeated-measures ANOVA with fixed factors PAIR, LEVEL and ΔF_0, and random factor SUBJECT. All fixed factors and interactions were significant ($p<0.0001$ for all but ΔF_0: $p=0.001$, and ΔF_0*LEVEL: $p=0.0024$). The SUBJECT factor was not significant, nor was its interaction with LEVEL, or with LEVEL*PAIR), but all other interactions were significant, reflecting differences between subjects in their detailed pattern of performance.

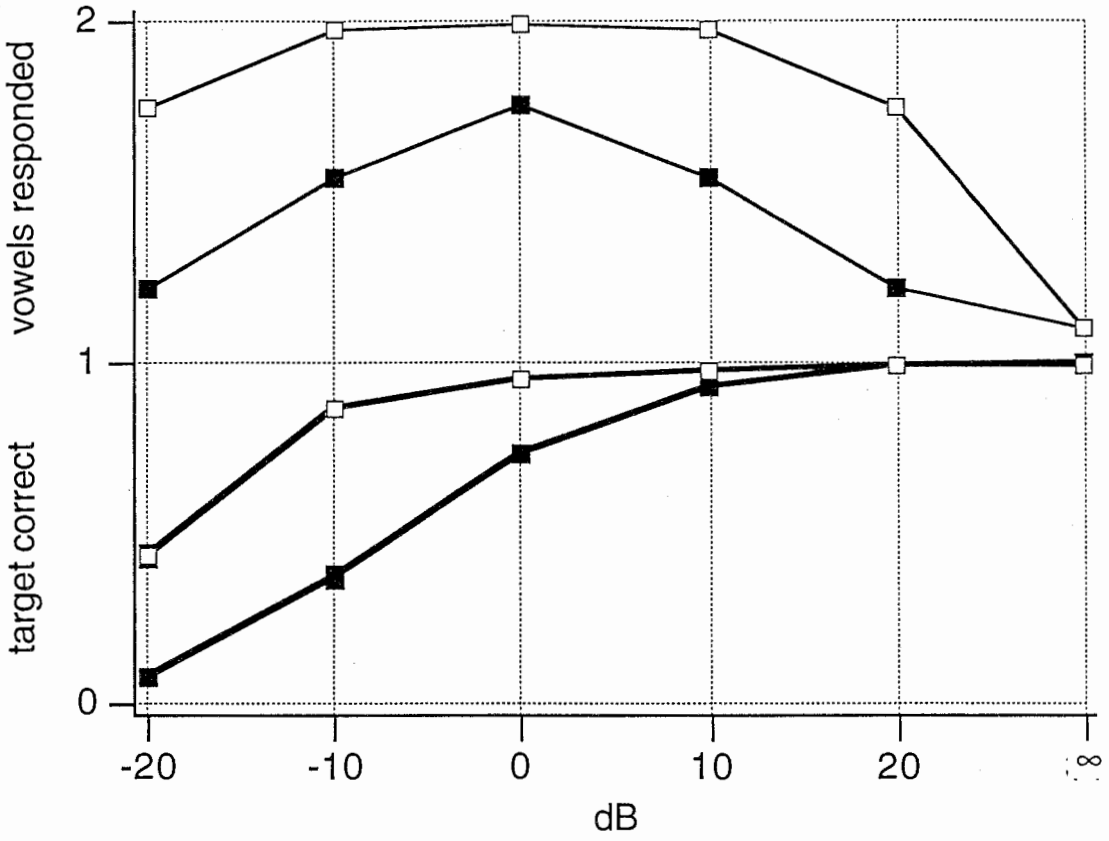


Fig 2. Top: number of vowels responded per stimulus, bottom: target identification rate. The abscissa is level of target vowel relative to background vowel. Filled symbols are for unison, open symbols are for a 6% difference in F_0. The rightmost point (∞) represents single vowels.

4.3.2.2. Effect of level

Fig. 2 (lower part) shows the average identification rate as a function of relative level between vowels, for both ΔF_0s. Identification increases monotonically in both cases. The upper part of Fig. 2 shows that the number of vowels responded is largest when both vowels have the same level, and drops off if either vowel is stronger.

A goal of Experiment 1 was to determine the best "operating point" for subsequent experiments. A performance level of 60-70% seems appropriate to avoid ceiling effects, and is not so low as to discourage subjects. Interpolating from the results, a target level of -15 dB seems appropriate for experiments that use 6% ΔF_0 as a baseline. If 0% ΔF_0 were the baseline, a target level of -5 dB might be better. If the direction of the effect is known, other choices may be preferable. For example a target level of -10 dB gave the largest ΔF_0 effects in our data.

Large effects provide no benefit if uncontrolled variability also becomes large. To check for this possibility we formed the ratio between the ΔF_0 effect (difference in scores between 6% and unison) averaged over all conditions other than level, and the standard deviation calculated over these conditions. The ratio was highest at -10 dB (Fig. 3).

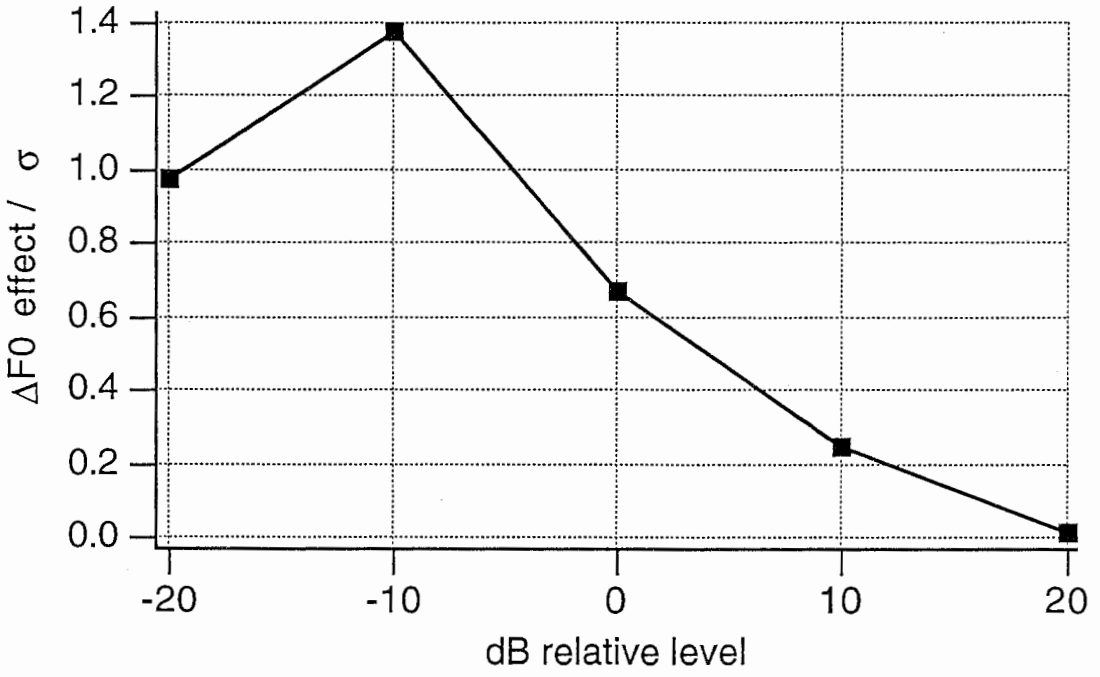


Fig. 3. Ratio between the ΔF_0 effect (difference between identification rates at 0 and 6% ΔF_0) and its standard deviation over all conditions (other than ΔF_0 and level).

4.3.2.2. Effect of ΔF_0

The size of the ΔF_0 effect depends on level, and is largest when the target vowel is -10 dB below the vowel it is mixed with. Interpolating between data points in Fig. 2 and taking the *horizontal* distance between curves at an ordinate of about 70%, yields a difference of about 14 dB, which is comparable to the 17 dB shift in masked threshold measured by Culling et al. (1994) in an adaptive task.

Combination-correct rates for the equal-level condition are plotted in Fig. 1 together with results of previous studies. Two aspects are striking. One is the relatively large size

of the ΔF_0 effect in our results, probably a benefit of the one-or-two response task. The other is the relatively high rate obtained at 6% ΔF_0, despite the fact that the one-or-two response task could have lead to relatively low combination-correct scores.

At a ΔF_0 of 6%, almost all stimuli evoke two-vowel responses, at unison the proportion is much smaller (Fig. 2, top). ΔF_0 thus functions as a strong "multiplicity" cue. When a subject only responds one vowel, the response to the other one is counted as false, so such "multiplicity" cues contribute to magnify effects on identification. A similar remark might be made concerning the threshold technique used by Summerfield (1992, Summerfield and Culling, 1992a, Culling et al., 1994): correct responses are impossible unless the interval containing the target stimulus is correctly recognized. This may also depend on "multiplicity" cues.

The ΔF_0 effect is marked even when the relative level of the target is low. This can be interpreted as evidence that the auditory system uses strategies other than harmonic enhancement: enhancement requires knowledge of the target F_0, which should be difficult to estimate at low SNR. In a similar experiment that also manipulated relative level and ΔF_0, McKeown (1992) instead found that ΔF_0 effects were reduced beyond 10-12 dB level mismatch. That may have been the result of a floor effect: the identification levels reported were overall much lower than the ones we found.

The dependency of identification on level with and without ΔF_0 (lower part of Fig. 2) is similar to the dependency of recognition rate on SNR in a speech recognition system, with and without noise-reduction processing based on cancellation (de Cheveigné, 1994; de Cheveigné, et al., 1994a).

4.3.2.3. Vowel pair, subject and session effects

Differences in patterns between vowel pairs and subjects are discussed in Appendices B and C, and serial differences across sessions in Appendix D.

4.4. Conclusion

The one-or-two response task improves the sensitivity of the double-vowel identification paradigm, probably by tapping "multiplicity cues" that the classic two-response task ignores. The number of vowels responded is in itself an interesting measure. Subjects find the task more "natural", and easier than when two responses are required.

Ceiling effects are reduced and effects are stronger if the target vowel level is reduced relative to the background. A level of -15 dB seems appropriate for experiments that use the 6 % ΔF_0 condition as a baseline.

5. Experiment 2

5.1. Introduction

Experiment 2 was designed to verify whether the ΔF_0 effect found in classic "double vowel" experiments depends on the phase patterns of the vowel stimuli. Such might be the case if segregation occurred according to either of two mechanisms that have recently been proposed: PPA (Pitch Period Asynchrony), and beats.

5.1.1. PPA

An F_0 difference is equivalent to a gradually increasing time shift of one wave form relative to another. If a vowel's short-term energy is not uniformly distributed within its period (given some definition of "short-time energy"), then the masking it causes or receives may vary with time alignment. ΔF_0 might in this way cause either vowel or both to be better perceived. This is the Pitch Period Asynchrony (PPA) mechanism.

Summerfield and Assmann (1991) investigated whether such a time lag per se is sufficient in the absence of mistuning. They presented subjects with synthetic vowels of same F_0 (50 or 100 Hz), with and without a time shift of one half a period. The time shift produced an significant improvement in identification at 50 Hz, but not at 100 Hz. Assmann and Summerfield (1994) did find a significant improvement at 100 Hz, as well as other evidence that PPA contributes to segregation. However they failed to replicate the time-shift effect with inexperienced subjects.

Estimates of the equivalent rectangular duration (ERD) of the auditory temporal window are of the same order (6-13 ms) (Plack and Moore, 1990) as the fundamental periods used in double-vowel experiments, so one might expect period features to be smoothed out too much for PPA to work. However Kohlrausch and Sander (1995) found that masking of a short pure-tone target varied by as much as 17 dB within the period of a 100 Hz masker. The variation was smaller (about 6 dB) at a fundamental of 220 Hz. At a given masker F_0 the variation was large when the masker components were in sine phase or m+ Schroeder phase (which both presumably produce highly modulated patterns of activity within auditory channels), but small with a m- Schroeder phase masker (which presumably produces flatter modulation patterns).

Several experiments suggest that vowel identification might depend on uneven masking within a masker's fundamental period. Moore and Alcántara (1995) synthesized harmonic "vowels" with a fundamental of 100 Hz and a spectral envelope that was flat on average. "Formants" were defined by amplitude modulation of groups of two consecutive harmonics at a rate of 10 Hz. For cosine phase the stimuli could be identified as vowels, despite their flat average spectrum. For random phase, identification was at chance level.

Traunmüller (1987) used the amplitude spectrum of a glottal source together with the phase spectrum of a glottal tract to synthesize nine Swedish "vowels". There were no spectral amplitude peaks present to signal the formants, but several subjects could label the stimuli consistently if the F_0 was low enough (71 or 100 Hz). Labeling was less consistent at higher frequencies (141 and 200 Hz), and at 283 Hz it fell to chance level. The "phase vowels" were intelligible via earphones, but not when presented through a loudspeaker in an ordinary room.

Palmer, et al. (1987) observed a change with phase of the position of the F1 phoneme boundary along a /e/-/I/ continuum. The harmonic manipulated was the 4th harmonic (500Hz) of a 125 Hz fundamental. The boundary moved down from 450Hz to 430 Hz when the phase shifted by 90 degrees relative to the phase produced by a Klatt synthesizer. This suggests that the phase shift produced a 20 Hz rise in the perceived F1 of the stimuli. The authors also performed a physiological experiment in which similar stimuli (with a fundamental of 100 Hz) were presented to guinea pigs, and the response was recorded from a population of auditory-nerve fibers. Without the phase shift, fiber responses below 1 kHz were equally dominated by frequencies of 400 or 500 Hz. With a

90 degree phase shift, they were dominated mainly by the higher component. Such a change in response pattern could explain a rise in perceived F1.

All three results can be explained in a similar fashion. Stimuli with cosine phase and a flat spectrum (as in Moore's experiment) have a peaked wave form that produces strongly modulated activity within peripheral channels, as long as the F_0 is low enough and the channel CF high enough (Horst, et al., 1986). Vowels in sine, cosine or Klatt phase also create relatively strong modulation within peripheral channels (see Fig. A-3). Within the dips of this modulation, masking may be relatively weak. Raising or lowering the level of a group of components, as in Moore's experiment, is equivalent to adding them to the original signal in the same or opposing phase. The added components stand out during the dips in the cosine masker, and are perceived as vowels. The random-phase masker has no such dips, hence the lack of effect in that case. In Palmer's experiment, the phase shift of the 500 Hz component can also be interpreted as the addition of this component to the original wave form (with suitable phase and amplitude), so that it stands out within the interval of low activity within the period, as suggested by the physiological recordings. If phase transitions at the formants in Traunmüller's stimuli produced temporal effects similar to local phase shifts, a similar explanation would account for his results.

PPA effects depend on particular phase-dependent wave form patterns, and vowel identification depends on phase in a variety of situations. One may ask if the ΔF_0 effects found in classic double-vowel experiments also depend on phase. In the extreme, one might wonder if they exist at all when vowels are synthesized with random phase patterns!

5.1.2. Beats

PPA requires the auditory system to have a temporal resolution fine enough to follow fluctuations on the scale of the fundamental period. However wave form interaction between vowels can also produce interference patterns that are static, or vary according to slow beat patterns. Whereas PPA requires particular *intra*-vowel phase patterns that produce "peaky" wave forms or patterns of activity within peripheral channels (together with an inter-vowel phase pattern equivalent to a time lag), beats essentially depend on on-going *inter*-vowel phase relationships.

Culling and Darwin (1994) suggested that beats in the low-frequency (F1) region might explain improvements of identification with small ΔF_0s. Assmann and Summerfield (1994) found that successive 50 ms intervals excised from a 200 ms double vowel were not equally identifiable. Identification rates for the best interval were compatible with the idea that the auditory system takes advantage of beats to choose, within the 200 ms stimulus, a favorable interval on which to base identification.

Two partials that fall within a peripheral filter channel will beat at a rate equal to their frequency difference. The beat will be appreciable if the partials are similar in amplitude. It will affect identification if it occurs in a spectral region that determines a vowel's identity, and at a rate that is not too fast to be tracked by the auditory system. Beats may affect not only the short-term spectrum, but also the overall spectrum of the stimulus. Unless all beat periods are integral divisors of the stimulus duration, the overall spectrum will depend on starting phases: the amplitude spectra of the constituent vowels do not suffice to determine the spectrum of the sum.

Fig. 4 illustrates the beat pattern that might occur for an /ae/ double vowel in which the /a/ is 12 dB more intense than the /e/ (so that the spectral envelopes have similar levels near formants F1 and F2 of /e/). The beating might effectively signal the presence of the /e/ despite the low spectrum level at its formants (Assmann and Summerfield, 1994).

If phase-dependent wave form interactions affect vowel identification and contribute to ΔF_0 effects, one may wonder whether the "ΔF_0 effect" is due in part to some particularly unfavorable inter-vowel phase pattern that produces low identification scores at unison. If so, the phenomenon of "F_0-guided segregation" might be specific to this phase pattern.

Experiment 2 measured the ΔF_0 effect in three conditions. In the SS condition both vowels were in sine phase. Wave forms were thus peaky (as in Klatt phase) and aligned at unison, providing ideal conditions for a PPA effect. In the RR condition, both vowels had the same "random" phase pattern. There was thus no clear peak within the period, but temporal features were nevertheless aligned at unison, and shifted when there was a ΔF_0 (possibly supporting a "weak" form of the PPA mechanism). In the RR' condition, each vowel had a different "random" phase pattern. Wave forms thus shared no particular temporal feature, and there was no particular alignment at unison, so even the "weak" form of PPA should be defeated. In both the SS and the RR conditions, inter-vowel phase was zero at unison so vector summation produced a spectrum equal to the sum of the spectra of the individual vowels. In the RR' condition the spectrum at unison was the result of "random" vector summation.

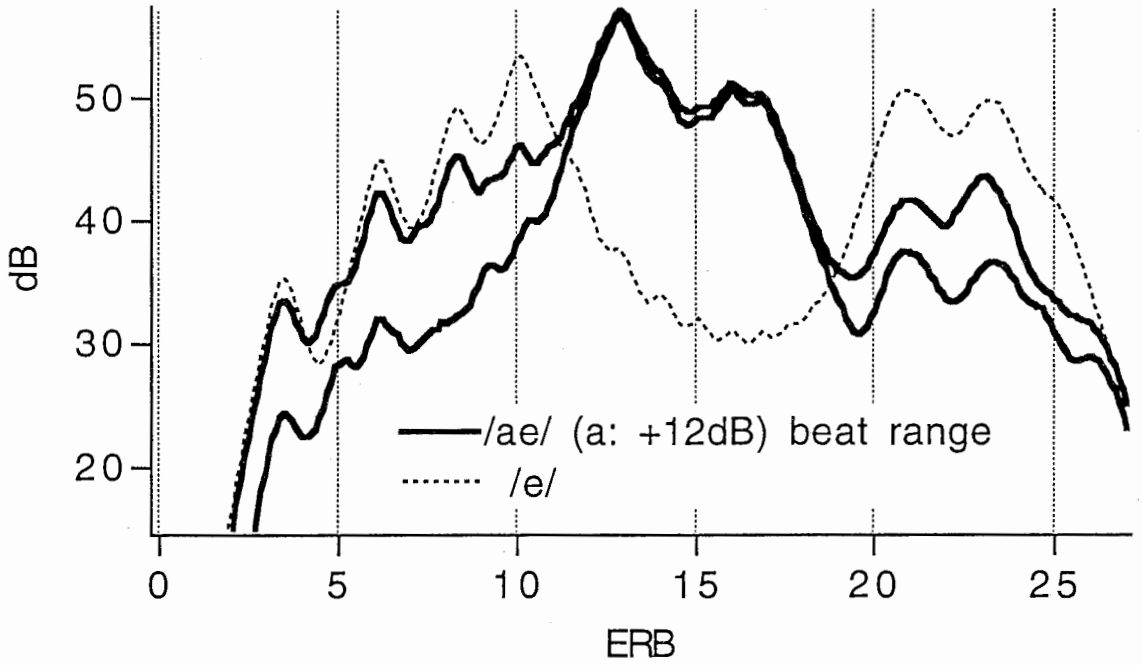


Fig. 4. Excitation patterns for an /ae/ double vowel in which the /a/ is 12 dB more intense than the /e/. The vowels had fundamentals of 124 and 132 Hz respectively. The excitation patterns were derived from an FFT based on a 16 ms Hanning-shaped window, smoothed according to the formulae of (Moore and Glasberg., 1983). The thick curves delimit the range of variation of the excitation pattern for the combined stimulus over its 250 ms duration. The thin dotted curve represents the excitation pattern for /e/ alone.

5.2. Methods

Single vowels were synthesized in sine phase (S) and either of two random phase patterns (R and R') at frequencies of 124 and 132 Hz, allowing ΔF_0s of 0% and 6.45% to be explored. See Appendix A. for details. Vowels were paired with an inter-vowel level mismatch of 15 dB. Phase patterns were either SS (both vowels in sine phase), RR (both vowels with the same "random" phase pattern) or RR' (different random phase patterns). The random phase patterns are those labeled 1 and 2 in Fig. A-3. There were (20 ordered pairs) x (3 phase conditions) x (2 ΔF_0s) x (2 F_0 orders) = 240 stimuli. These were

interleaved together with stimuli of Experiments 3 and 4 and single vowels (200) in blocks of 600 stimuli.

5.3. Results

Results are displayed in Fig. 5. The ΔF_0 effect is practically identical for all phase patterns. There are strong differences between subjects (crosses in Fig. C-1) that can partly be predicted from individual results in Exp. 1.

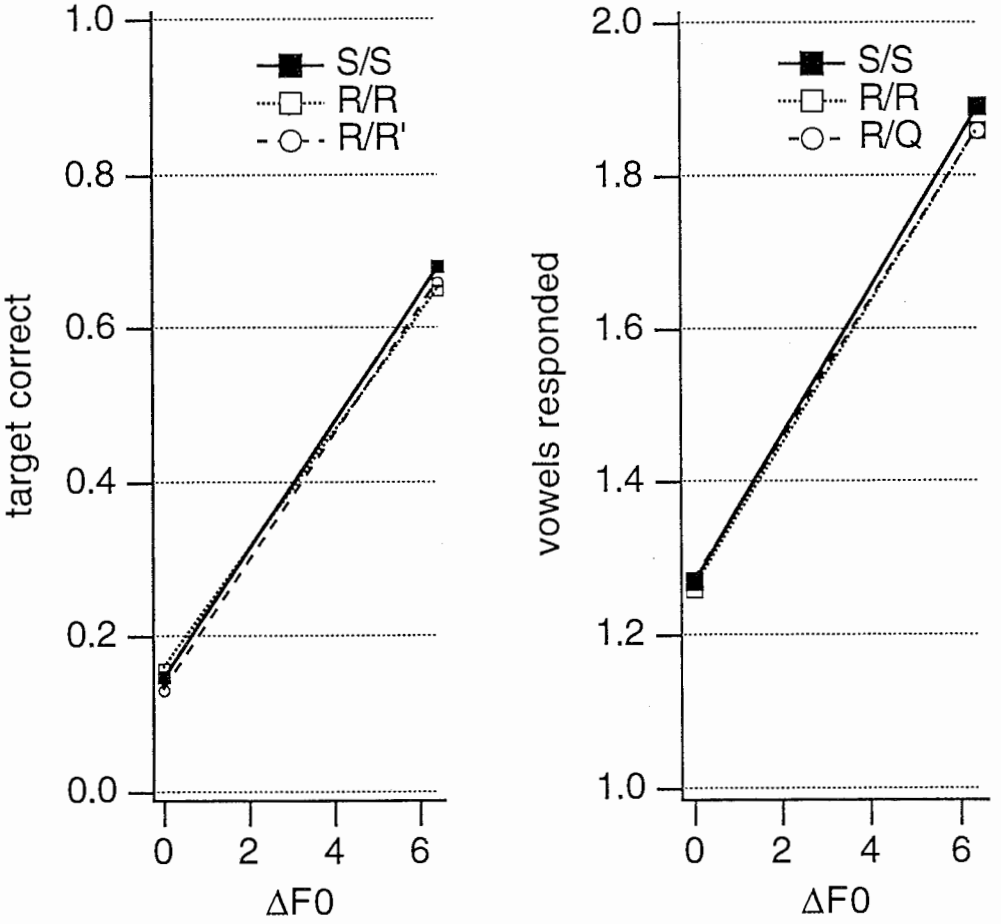


Fig. 5. Identification rate (left) and number of vowels responded (right), as a function of ΔF_0, for three phase patterns.

5.4. Conclusion

The classic ΔF_0 effect is in no way specific to Klatt or sine intra-vowel phase patterns that produce peaked wave forms, or to a particular inter-vowel phase pattern. Either our phase manipulations failed to affect the cues that underlie PPA or beat mechanisms, or else these mechanisms are not responsible for segregation with ΔF_0.

6. Experiment 3

6.1. Introduction

In a previous experiment (de Cheveigné et al., 1995), we presented subjects with pairs of vowels, each of which was either harmonic or inharmonic. We found that vowels were better identified if they were inharmonic than if they were harmonic. They were also better identified if the background was *harmonic* rather than inharmonic. These results were interpreted as supporting a particular class of segregation mechanism: harmonic cancellation.

However, all our stimuli were synthesized with an initial sine phase. Harmonic stimuli kept this phase throughout the stimulus, but inharmonic stimuli could be interpreted as gradually shifting to a random phase pattern. If phase affected identification, then the effects we attributed to harmonicity might simply have been an artifact due to phase. If so, harmonic stimuli synthesized with those phase patterns should show a similar pattern of effects.

Specifically, given two harmonic vowels synthesized with a ΔF_0 sufficient to cause segregation (3% in our previous experiment, 6.45% in the present experiment), if we find the following pattern:

$$R/S > R/R'$$
$$S/S > S/R$$
$$R/S > S/S$$
$$R/R' > S/R$$

where X/Y represents the identification rate of a target in state X with a background vowel in state Y, then we will have identified a possible artifact in the results of de Cheveigné et al (1995).

6.2. Methods

Conditions S/S and R/R' were shared with Exp. 2, others were interleaved with conditions of that experiment. In all conditions, ΔF_0 was 6.45%.

6.3. Results

Identification rates were calculated for the four conditions of interest, averaged over F_0 and session, and transformed according to the formula arcsin(2*rate-1) to obtain distributions closer to normal. . Each data point was based on 10 judgments. Data were submitted to a repeated measures ANOVA with fixed factors PAIR and PHASE, and random factor SUBJECT. Neither PHASE nor any interaction involving it was significant, implying that the pattern of phase effects (supposing they exist) is too variable across subjects to be generalized to the population. To test whether PHASE effects exist for *any* individual subjects, more sensitive tests were performed. Data were averaged over SUBJECT rather than SESSION, and submitted to a *fixed* effect ANOVA with factors PHASE and PAIR. No effect of PHASE or interaction was significant. Finally data were averaged over PAIR instead of SUBJECT and analyzed in the same fashion , with the same negative result.

Data from the three phase conditions (S/S, R/R, R/R') at unison in Exp. 2 were also averaged over session and analyzed for phase effects. A repeated measures ANOVA with fixed factors PHASE and PAIR and random factor SUBJECT showed no significant effect involving PHASE. Again, more sensitive analyses were performed:

1) Data were averaged over SUBJECT rather than SESSION and submitted to a repeated-measures ANOVA with factors PHASE and PAIR, and random factor SESSION. The PHASE effect and its interaction with PAIR were significant (respectively $p=0.01$ and $p<0.0001$).

2) Data were averaged over PAIR and submitted to a repeated measures ANOVA with fixed factors PHASE and SUBJECT, and random factor SESSION. PHASE and its interaction with the SUBJECT factor were significant (respectively $p=0.01$ and $p<0.0001$).

The small phase effect at unison is explainable from the fact that phase affects the outcome of vector summation, and therefore the overall spectra of the stimuli.

6.4. Conclusion

There is little evidence of any phase effects, except at unison where phase determines the spectrum as a result of vector summation of the single vowel components. Even in that case the effects are small. There is no certainly no evidence of the artifact hypothesized in the Introduction.

7. Experiment 4

7.1. Introduction

Experiment 3 ruled out the possibility of a phase artifact in the experiment reported by de Cheveigné et al. (1995). We nevertheless wished to replicate the main conditions of that experiment to lift any doubts about its generality. Experiment 4 repeated three crucial conditions (H/H, I/H and H/I at 6.45% ΔF_0) using stimuli designed to minimize the usefulness of PPA or beat cues.

From past results we can assume that segregation occurs for harmonic vowels at this ΔF_0. Perturbing the harmonicity of the target should impair the efficacy of harmonic enhancement mechanisms. Likewise, making the background inharmonic should disrupt harmonic cancellation.

7.2. Methods

One condition (H/H) was common with Experiments 2 and 3 (R/R' phase, 6.45% ΔF_0). The other two (I/H and H/I) were interleaved with conditions of those experiments. Steps were taken to reduce the usefulness of PPA or beat cues:

1) Intra-vowel starting phase was "random" to reduce the salience of temporal features within the period. Phase patterns were different between vowels, so any residual temporal features were not common to both vowels.

2) Inter-vowel starting phase was "random", and remained "random" with the ongoing phase shifts caused by ΔF_0 or inharmonicity. There is thus little reason to expect wave form interactions to favor one condition over another.

3) All component frequencies were multiples of 4 Hz, so the true period of all stimuli was 250 ms, the effective duration of the stimulus. The long-term spectrum of the stimulus thus could not depend on the choice of starting phases.

4) Within pairs containing an inharmonic vowel, no partials were closer than 16 Hz. To use spectral changes caused by beats, the auditory system must sample the beat pattern with a resolution better than about 30 ms. Of course, this cannot be excluded, but we expect it to be more difficult than with slower beats. See Appendix A for details.

7.3. Results

7.3.1. Double vowels

Identification rates were averaged over frequency and session, and transformed according to the formula arcsin(2*rate-1). A repeated measures ANOVA was performed with fixed factors HARMONICITY and PAIR, and random factor SUBJECT. All factors and interactions were significant ($p<0.0001$), except HARMONICITY*SUBJECT. The effects are illustrated in Fig. 6. Identification is better by 21% when the ground vowel is harmonic than when it is inharmonic. This is evidently not due to exploitation of a "multiplicity cue": the number of vowels responded is nearly the same. Harmonicity of the target makes no significant difference to either identification or response count.

The effect of ground harmonicity is consistent with what we found previously (de Cheveigné et al., 1995), but seven times larger. This confirms once again the hypothesis of harmonic cancellation. On the other hand the lack of effect of target harmonicity contrasts with our previous finding of a 3% advantage for inharmonic targets. An explanation we had offered for that result (paradoxical because opposite the prediction for harmonic enhancement), was that the auditory system applied cancellation indiscriminately, and that a harmonic target might fall victim to it more easily than an inharmonic target. The relatively low level of the target level here (-15 dB) may make its cancellation unlikely. Cancellation requires estimation of the F_0 of the vowel to cancel, which is difficult to at low SNR.

It is somewhat surprising that inharmonic targets don't give *worse* recognition scores, as subjects report inharmonic stimuli as somewhat strange and not vowel-like.

Single vowels were identified better than 99.5% whatever their harmonic state, but it is interesting to note that inharmonic vowels evoked more responses (63%) than harmonic vowels (8%). Inharmonicity seems to function as a "multiplicity" cue similar to an F_0 difference between harmonic vowels.

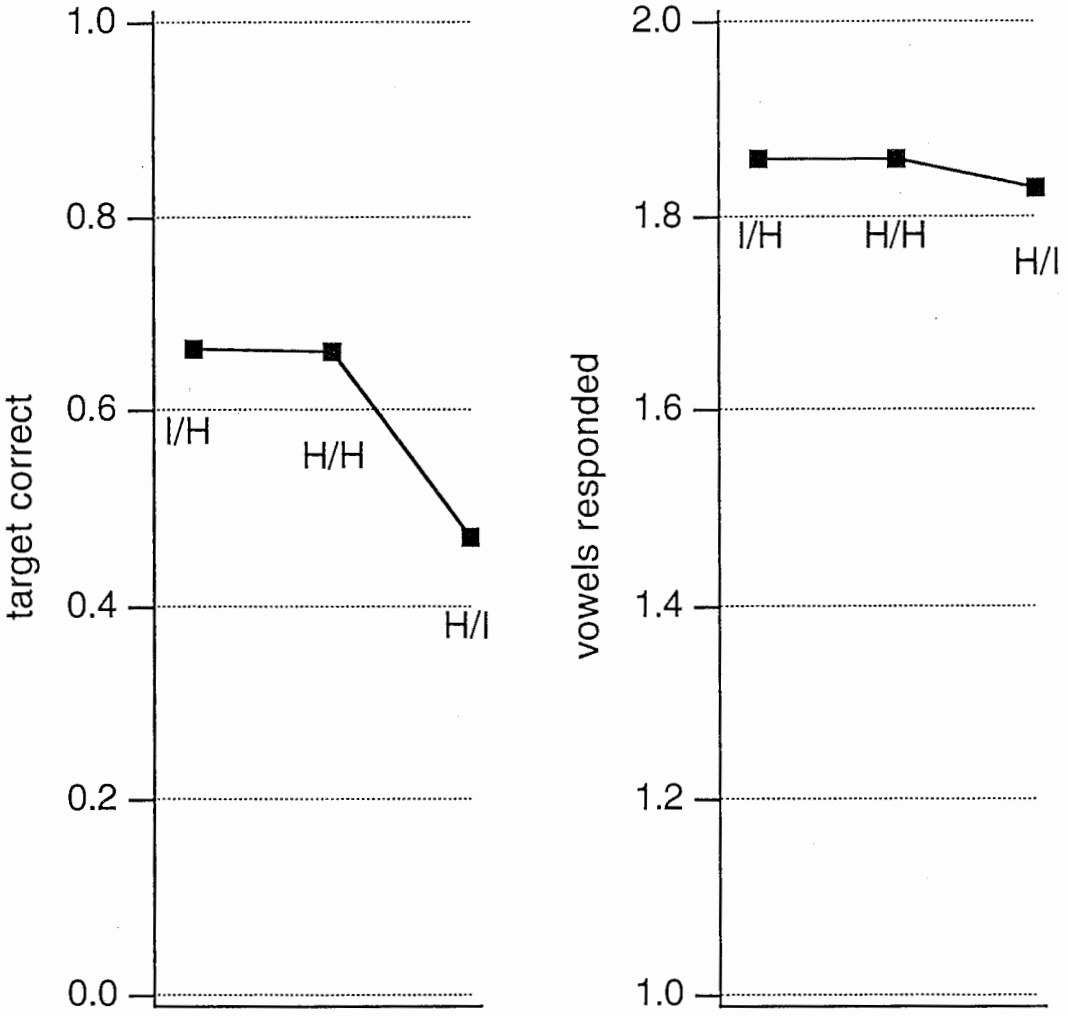


Fig. 6. Identification rate (left) and number of vowels responded (right) as a function of target/ground harmonic state. Nominal ΔF_0 is 6.45%.

7.4. Conclusion

We successfully replicated our previous finding that targets are better identified when the ground is harmonic. Given our precautions to avoid PPA and beat cues, and the lack of effect of phase, we can rule out the possibility of a phase-related artifact in the results of de Cheveigné et al. (1995). On the other hand we failed to replicate our previous (paradoxical) finding of a better identification of inharmonic targets.

8. Experiment 5

8.1. Introduction

Our methods appear to yield larger effects than usually reported in double-vowel experiments. This improvement can be attributed to two factors: the one-or-two-response task, and the level mismatch. Experiment 5 seeked to determine the part of each, by repeating some conditions of the previous experiments with a classic two-vowel forced-response task, using the same subjects and stimuli. It also gave us a basis for comparison with previous studies.

The two-response task reduces the usefulness of "multiplicity cues", so identification depends essentially on cues that determine the efficacy of "unmasking" mechanisms.

8.2. Methods

We replicated four conditions of Experiments 2 and 4: H/H at unison, and H/H, I/H and H/I at a ΔF_0 of 6%. These were interleaved with conditions of Experiments 6 and 7, in stimulus sets containing 400 double vowels in random order, and no single vowels. Subjects were five (K, M, N, T, U,) of the six that participated in the previous experiments. They were informed that all stimuli contained two vowels, and that they must respond a pair of two different vowels. The "x" response was no longer allowed. They were told to make their "best guess" if they could not hear two vowels.

Stimuli were presented in five sessions, on different days. Due to a mistake, two conditions (I/H and H/I at unison) were not included in the stimulus set on the first session. Statistical analyses were conducted on either all five sessions excluding those conditions, or the four last sessions including all conditions, as appropriate.

8.3. Results

Fig. 7 compares the average results of the five subjects with those they obtained for the same conditions in Experiments 3 and 4. Identification was overall better with the two-vowel forced response task, as might be expected, since subjects can no longer get away with answering only one vowel. Improvement was greatest for conditions that gave low rates, leading to a reduction in effect size that was appreciable for the ΔF_0 effect, but small for the background harmonicity effect.

We expected rather large training effects over sessions with the two-vowel forced response task. This was hardly the case (see Appendix E).

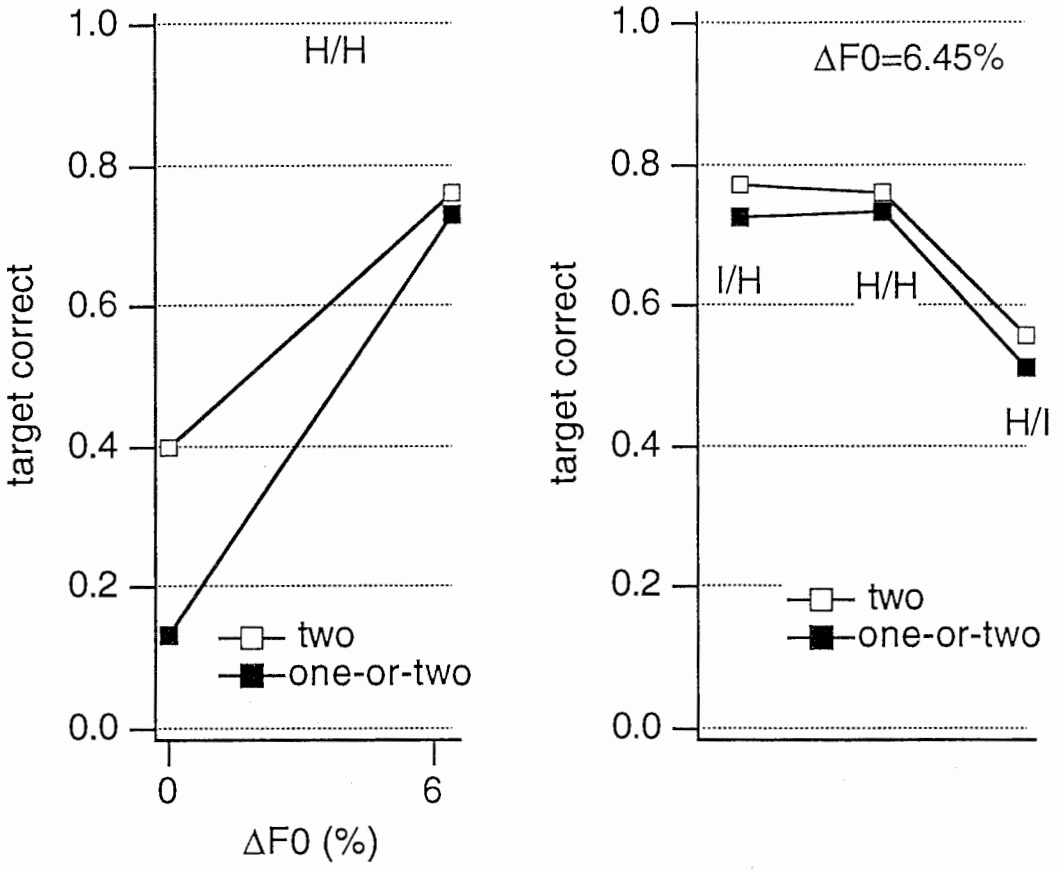


Fig. 7. Left: identification rate as a function of ΔF_0 for the same subjects using two different tasks (two-response and one-or-two response). Right: identification rate as a function of the harmonic states of target and ground for the same subjects using both tasks. Nominal F_0 difference is 6.45%,

8.4. Conclusion

The two-response task yielded higher identification rates than the one-or-two response task for the condition that lacked "multiplicity" cues (H/H at unison). It thus contributed to enlarge the ΔF_0 effect. It did not change the pattern of identification rates for conditions that already evoked multiple responses, apart from a slight uniform increase in identification rate.

9. Experiment 6

9.1. Introduction

Experiment 6 is a replication of all conditions of the experiment described by de Cheveigné et al. (1995), using the classic two-response task. Main differences with that experiment are a larger nominal ΔF_0 (6.45% rather than 2.9%), a lower relative level of the target (-15 dB rather than 0 dB), and different initial phase relationships (random rather than sine). Other differences concern the vowel set (one allophone each of five Japanese vowels, rather than ten allophones each of five French vowels), the definition of inharmonic vowels (see Appendix A), and the subjects (5 Japanese speakers rather than 30 French speakers).

9.2. Methods

Four of the conditions were those used in Experiment 5. Others were interleaved with those of Experiment 5 and 7. The task was the classic two-vowel response task. Inharmonic component frequency patterns were designed so that components of a double vowel were spaced at least 16 Hz apart (Appendix A). At unison in the I/I condition the vowels in a pair had different inharmonic patterns.

9.3. Results

Fig. 8 shows the results, together with those of de Cheveigné et al. (1995). The much larger effects in our new experiment tend to mask the similarity of ground harmonicity effect: targets are easier to identify with a harmonic background, except when the target is also harmonic and has the same F_0 (in which case harmonicity is of no avail). A major difference is that we no longer see the effect of target harmonicity found previously.

Several factors may explain the difference in effect size between the two experiments:
1) In the experiment of de Cheveigné et al. (1995), inharmonic sounds had partials displaced from harmonic values by random amounts smaller than 3%. In the present experiment, displacements were larger (up to 6.45%, with a minimum of 16Hz).
2) The -15dB target level avoided ceiling effects in the present experiment. In the previous experiment, intra-stimulus variability was supposed to play that same role by lowering overall identification rate. However the effectiveness of that measure may have been limited, for example if some allophones were identified perfectly whereas others produced systematic errors.

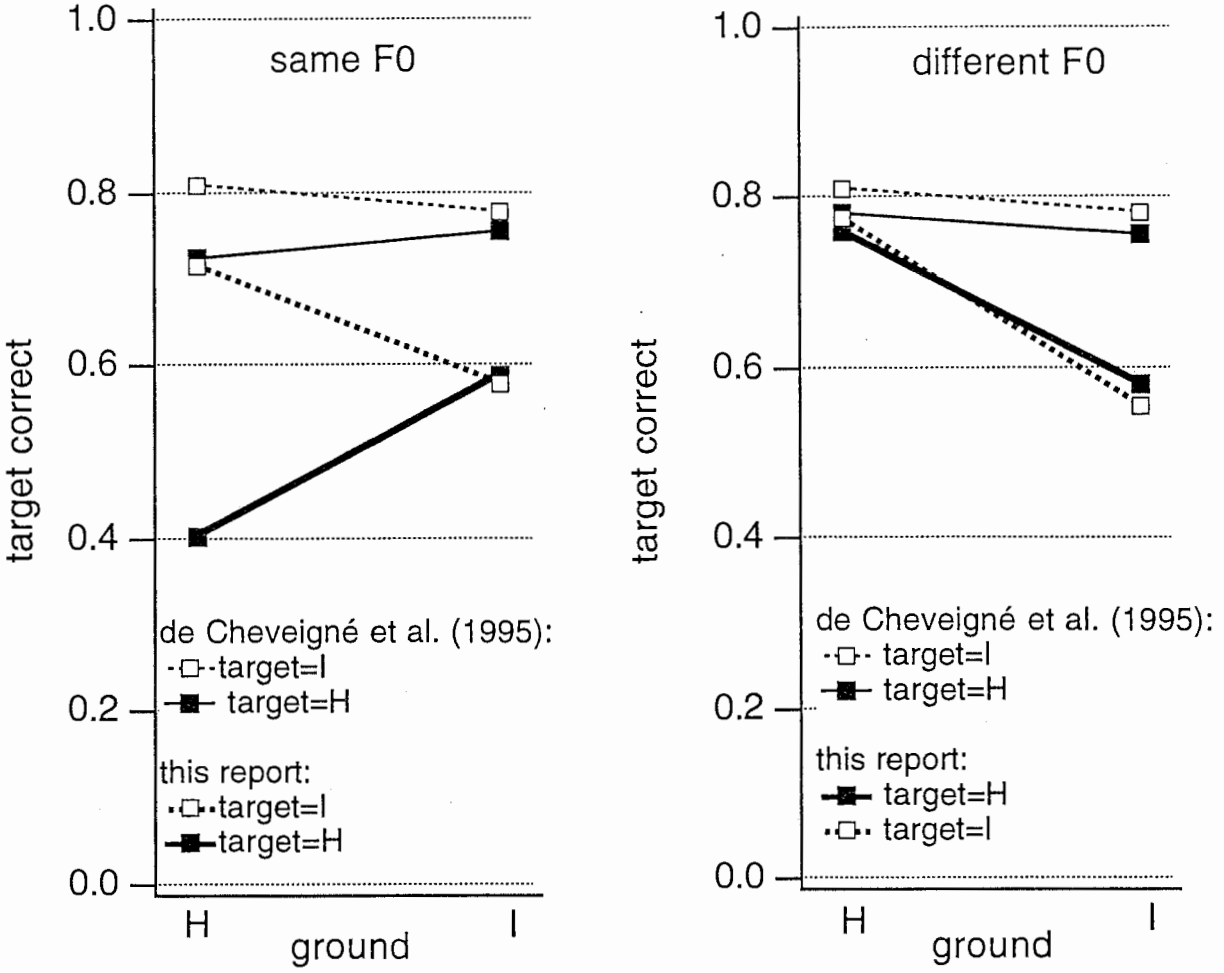


Fig. 8. Identification rate as a function of ground harmonicity for each of the target harmonicity states, at unison and with different F0s. Thin lines: (de Cheveigné, et al., 1995), thick lines: this study.

9.4. Conclusion

As we found before, identification of the target depends on the harmonicity of the *ground*. This is compatible with the harmonic cancellation hypothesis. Harmonicity of the target itself makes no difference, contrary to our previous paradoxical finding that harmonicity of a target made it *less* easy to identify, opposite to the predictions of harmonic enhancement. In either case, the data fail to support the hypothesis of harmonic enhancement.

10. Experiment 7

10.1. Introduction

In the I/I condition at unison in the previous experiment, the pattern of partial frequencies of each vowel in a pair was different. Experiment 7 compares that condition (denoted here I/I_diff) with a similar condition (denoted I/I_same) in which both vowels are inharmonic but with the *same* pattern of partial frequencies.

10.2. Methods

The I/I_same condition was interleaved with conditions of Experiments 5 and 6.

10.3. Results

Fig. 9 compares the I/I_same condition (inharmonic, same pattern) with the H/H (harmonic, same pattern) and I/_diff (inharmonic, different pattern) conditions. Identification in the I/I_same condition is similar to that in the H/H condition at unison. This can be due to the fact that partial frequencies are all the same (thus eliminating a "partial mismatch" cue), or to the fact that both conditions defeat any mechanism based on harmonicity.

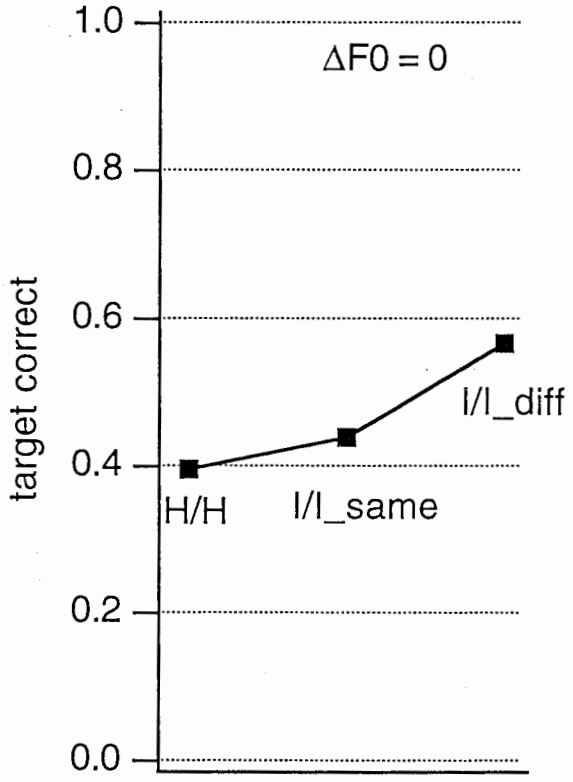


Fig. 9. Identification rate for several patterns of partial frequencies. Nominal F_0s of both vowels are the same.

10.4. Conclusion

Better identification in the I/I_diff condition relative to the other two could be due to increased mismatch between partials, or to a mechanism exploiting residual harmonicity within the inharmonic vowels. Inharmonic patterns are derived from harmonic patterns by a relatively mild perturbation. Inharmonic vowels have a clear pitch (which, interestingly, varies with the vowel). It is conceivable that segregation can occur based on residual

harmonicity, imperfect or local to a frequency region. The result of this particular experiment does not allow us to decide between these different possibilities.

11. General conclusions and summary

1) Reducing the level of one vowel relative to the other in the double-vowel identification paradigm improved sensitivity by avoiding ceiling effects.

2) Allowing subjects to respond *one or two* vowels instead of forcing them to respond two also contributed to improve the sensitivity of the paradigm. The number of vowels answered is a sensitive indicator of "multiplicity" cues. Subjects report that the task is easier, and training effects may be reduced.

3) Phase relationships among partials of a vowel or between vowels had no effect on vowel identification. We found no evidence to suggest a phase-related artifact in a previous experiment on harmonicity (de Cheveigné et al. 1995).

4) In a replication of that experiment, employing pairs of vowels that were either harmonic or inharmonic, target vowels were better identified when the background was harmonic, in agreement with the hypothesis of harmonic cancellation. Target harmonicity made no difference, contrary to the predictions of the harmonic enhancement hypothesis. We did not replicate our earlier finding of a better identification of *inharmonic* targets.

5) Results were similar when the one-or-two-response task was replaced by a classic two-response task, but effect sizes were reduced.

6) Overall, the results suggest that the auditory system uses the strategy of harmonic cancellation to segregate harmonic sounds such as vowels. It does not seem to make use of harmonic enhancement, and there was little evidence that beats or Pitch Period Asynchrony (PPA) are involved in segregation.

Acknowledgments

This research was conducted within the framework of a collaboration agreement between ATR Human Information Processing Laboratories and the Centre National de la Recherche Scientifique. I thank ATR for its kind hospitality, and the CNRS for leave of absence.

Cécile Marin, Jean Laroche and Steve McAdams participated in the preparation of these experiments. Hideki Kawahara, Minoru Tsuzaki, Kiyoaki Aikawa, Hiroaki Kato and Ikuyo Masuda contributed useful ideas and advice, and Rieko Kubo supervised the experiments. Particular thanks are due to Ikuyo Masuda and Kayoko Nakagawa who volunteered as subjects, and to the four paid subjects who each patiently responded to over 9000 stimuli. Thanks also to John Culling of the Nottingham Institute of Hearing Research for providing the software for stimulus synthesis.

Bibliography

Assmann, P. F. and Summerfield, Q. (1994). "The contribution of waveform interactions to the perception of concurrent vowels." J. Acoust. Soc. Am., 95, 471-484.

Bregman, A. S. (1990). "Auditory scene analysis". Cambridge, Mass.: MIT Press.

Carlyon, R.P. (1994). "Detecting pitch-pulse asynchronies and differences in fundamental frequency", J. Acoust. Soc. Am. 95, 968-979.

Culling, J. and Summerfield, Q. (1995). "The role of frequency modulation in the perceptual segregation of concurrent vowels." J. Acoust. Soc. Am., accepted for publication.

Culling, J. F. and Darwin, C. J. (1994). "Perceptual and computational separation of simultaneous vowels: Cues arising from low frequency beating." J. Acoust. Soc. Am., 95, 1559-1569.

Culling, J. F., Summerfield, Q. and Marshall, D. H. (1994). "Effects of simulated reverberation on the use of binaural cues and fundamental frequency differences for separating concurrent vowels." Speech Comm., 14, 71-95.

de Cheveigné, A. (1993a). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing." J. Acoust. Soc. Am., 93, 3271-3290.

de Cheveigné, A. (1993b). "Time-domain comb filtering for speech separation" (TR-H-016): ATR Human Information Processing Laboratories.
de Cheveigné, A. (1994,). "Strategies for voice separation based on harmonicity". Proc. ICSLP, Yokohama, 1071-1074.
de Cheveigné, A., Kawahara, H., Aikawa, K. and Lea, A. (1994a). "Speech separation for speech recognition." Journal de Physique IV, 4, C5-545-C5-548.
de Cheveigné, A., McAdams, S., Laroche, J. and Rosenberg, M. (1994b). "Identification de voyelles simultanées harmoniques et inharmoniques." Journal de Physique IV, 4, C5-553-C5-556.
de Cheveigné, A., McAdams, S., Laroche, J. and Rosenberg, M. (1995). "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement." J. Acoust. Soc. Am., to appear June 1995.
Hirahara, T. and Kato, H. (1992). "The effect of F_0 on vowel identification". In Y. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka (Eds.), "Speech perception, production and linguistic structure", Tokyo: Ohmsha, 89-112.
Holdsworth, J., Nimmo-Smith, I., Patterson, R. D. and Rice, P. (1988). "Implementing a GammaTone filter bank" (SVOS final report, annex C.): MRC Applied Psychology Unit.
Horst, J. W., E., J. and Farley, G. R. (1986). "Coding of spectral fine structure in the auditory nerve. I. Fourier analysis of period and interspike interval interval histograms." J. Acoust. Soc. Am., 79, 398-416.
Klatt, D. H. (1980). "Software for a cascade/parallel formant synthesizer." J. Acoust. Soc. Am., 67, 838-844.
Kohlrausch, A. and Sander, A. (1995). "Phase effects in masking related to dispersion in the inner ear. II Masking period patterns of short targets." J. Acoust. Soc. Am., 97, 1817-1829.
Lea, A. (1992). "Auditory models of vowel perception.", unpublished doctoral thesis, Nottingham.
Lea, A. P. and Summerfield, Q. (1992). "Monaural segregation of competing voices." Proc. ASJ committee on Hearing, H-92-31, 1-7.
McKeown, J.D. (1992). "Perception of concurrent vowels: the effect of varying their relative level", Speech Comm. 11, 1-13.
Meddis, R. and Hewitt, M. J. (1992). "Modeling the identification of concurrent vowels with different fundamental frequencies." J. Acoust. Soc. Am., 91, 233-245.
Moore, B. C. J. and Alcántara, J. I. (1995). "Identification of flat-spectrum vowels on the basis of amplitude modulation." J. Acoust. Soc. Am., 97, 3274.
Moore, B. C. J. and Glasberg, B. R. (1983). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns." J. Acoust. Soc. Am., 74, 750-753.
Palmer, A. R., Winter, I. M., Gardner, R. B. and Darwin, C. J. (1987). "Changes in the phonemic quality and neural representation of a vowel by alteration of the relative phase of harmonics near F1". In M. E. H. Schouten (Ed.), "The psychophysics of speech perception", (pp. 371-376). Dordrecht: Martinus Nijhoff.
Plack, C. J. and Moore, B. C. J. (1990). "Temporal window shape as a function of frequency and level." J. Acoust. Soc. Am., 87, 2178-2187.
Summerfield, Q. (1992). "Roles of harmonicity and coherent frequency modulation in auditory grouping". In M. E. H. Schouten (Ed.), "The auditory processing of speech: from sounds to words", Berlin: Mouton de Gruyter, 157-166.
Summerfield, Q. and Assmann, P. F. (1991). "Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony." J. Acoust. Soc. Am., 89, 1364-1377.
Summerfield, Q. and Culling, J. F. (1992a). "Auditory segregation of competing voices: absence of effects of FM or AM coherence." Phil. Trans. R. Soc. Lond. B, 336, 357-366.

Summerfield, Q. and Culling, J. F. (1992b,). "Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency". 124th meeting of the ASA.
Traunmüller, H. (1987). "Phase vowels". In M. E. H. Schouten (Ed.), "The psychophysics of speech perception", Dordrecht: Martinus Nijhoff, 377-384.

Appendix A. Stimuli.

Tokens of the five Japanese vowels /a/, /i/, /u/, /e/, /o/ were synthesized with spectral envelopes calculated according to the formulae specified by Klatt (1980). Formant frequencies and bandwidths are given in Table A-1. The first four formant frequencies have values suggested by Hirahara and Kato (1992), the fifth formant was set to 4200 Hz for all vowels. Formant bandwidths were given fixed values for all vowels, as used for example by Culling and Summerfield (1995). Wave forms of all stimuli were scaled to a uniform level of 65 dB RMS. before presentation. Table A-1 indicates the RMS. level of vowel wave forms *before* scaling (so-called "equal effort"), and also the dB(A) sound pressure levels for single vowels, measured with an artificial ear.

	/a/	/i/	/u/	/e/	/o/	BW
F1	750	281	312	469	468	90
F2	1187	2281	1219	2031	781	110
F3	2595	3187	2469	2687	2656	170
F4	3781	3781	3406	3375	3281	250
F5	4200	4200	4200	4200	4200	300
dB RMS. after synthesis	46.9	40.6	40.4	41.8	44.5	
dB(A) SPL	70.0	63.0	63.6	67.4	66.2	

Table 1: Formant frequencies and bandwidths of all synthetic vowels. Also shown are the RMS. levels in dB (re: 1.0) of the vowels after synthesis and before scaling to a uniform RMS. level, and the SPL levels in dB(A) produced by the scaled vowels, as measured with the artificial ear.

Fig. A-1 shows the spectral envelopes (scaled by the same amount as the wave forms). For each vowel pair, Table A-2 gives the level of the envelope at formants F1 and F2 of the first vowel, relative to the envelope level of the second vowel. This indicates the degree to which the formants of the target "stick out" of the envelope of the background vowel. Fig. A-2 shows estimates of the excitation patterns for each vowel. Excitation patterns were calculated by taking the FFT of a 16 ms Hanning-shaped window (2 periods) of a 100 Hz vowel, and applying spectral smoothing according to formulae of Moore and Glasberg (1983).

	F1	F2
ai	37	37
au	31	14
ae	25	27
ao	11	5
ia	20	21
iu	2	25
ie	12	5
io	14	36
ua	21	-9
ui	6	25
ue	13	15
uo	14	25
ea	19	31
ei	24	27
eu	19	38
eo	-1	47
oa	20	-6
oi	24	28
ou	20	22
oe	1	17

Table A-2. For each vowel pair, the table gives the spectral envelope level of the first vowel at formants F1 and F2, relative to the envelope of the second vowel (vertical distance between curves in Fig. A-1).

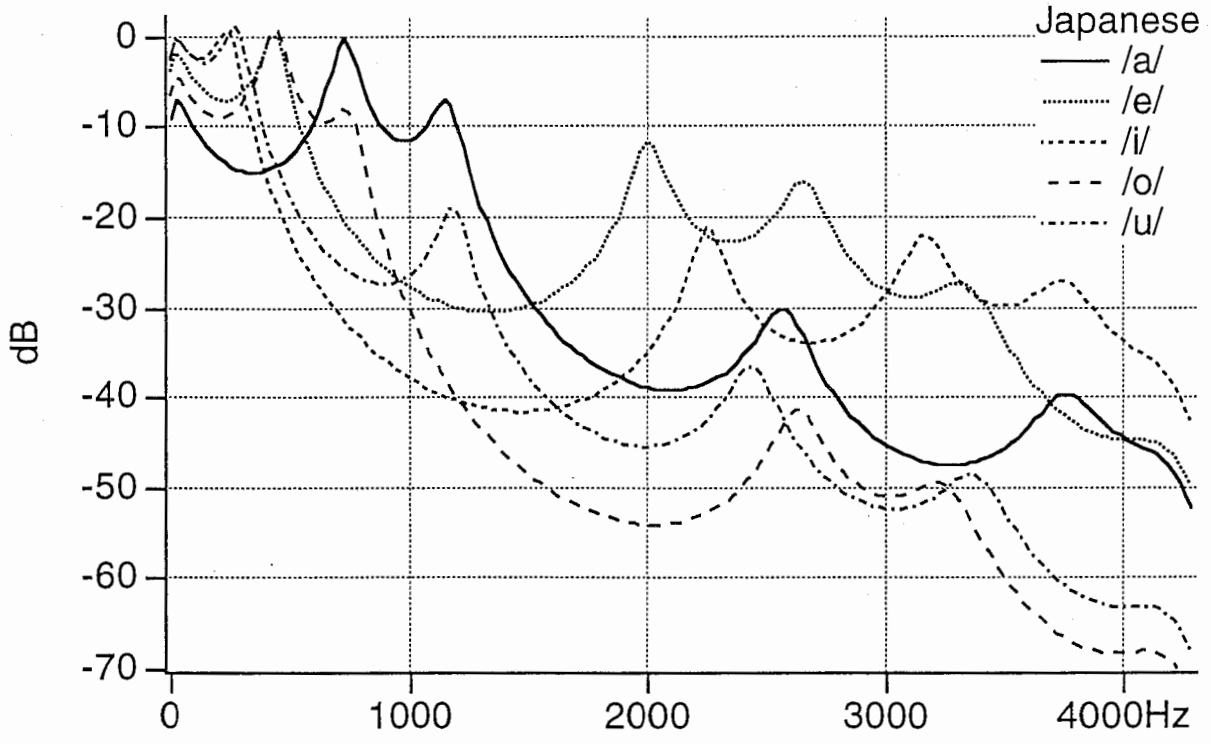


Fig. A-1: Envelopes of the synthetic vowels.

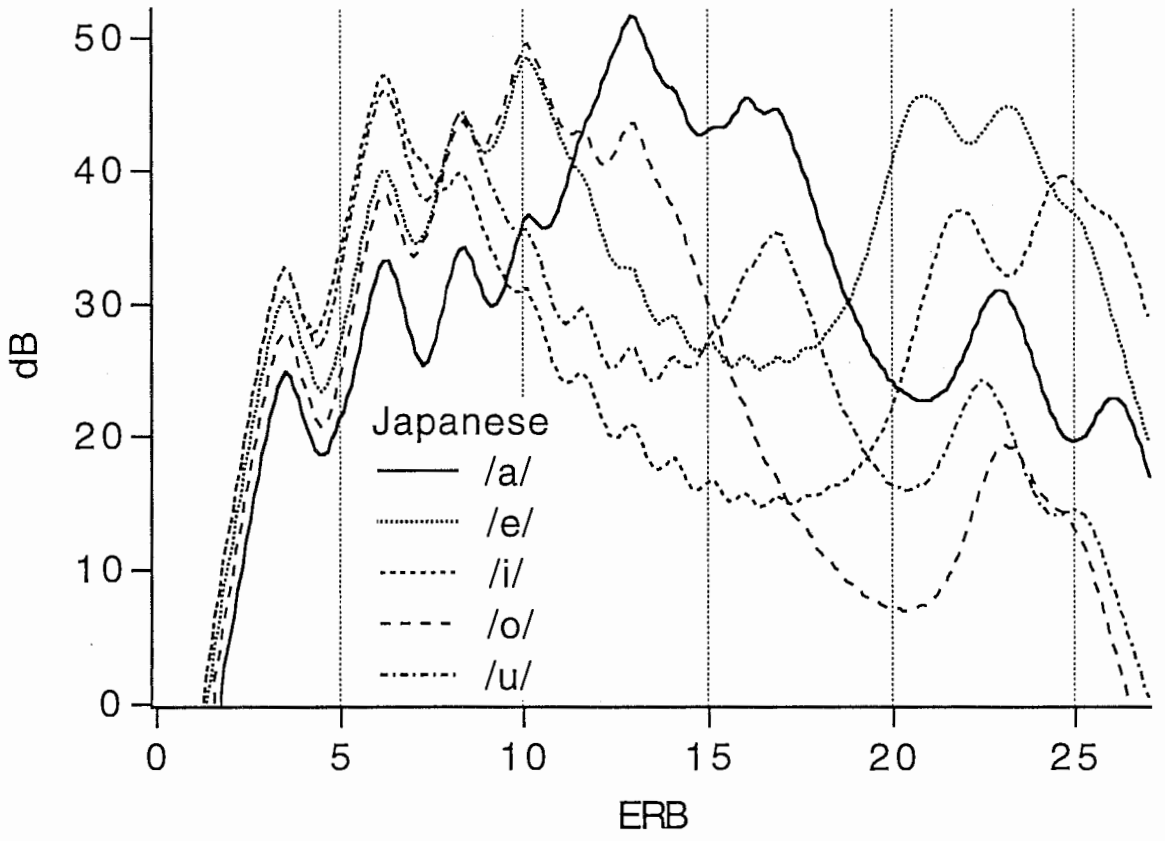


Fig. A-2: Excitation patterns calculated for the synthetic vowels.

Stimuli for Experiment 1 were synthesized in "Klatt phase" with F0s of 125 and 132.5 Hz. Duration was 200 ms, including 20 ms raised-cosine onset and offset ramps.

Stimuli for Experiments 2-7 were synthesized in either sine phase (S) or one of two "random" phases (R and R'), with F0s of 124 and 132 Hz and a duration of 270 ms (including 20 ms raised-cosine onset and offset ramps). The "effective" duration of 250 ms is the period of 4 Hz which divides the frequencies of all stimulus components. All beats frequencies are multiples of 4 Hz and produce an integer number of periods within the stimulus.

The "random" phase patterns were selected among ten random patterns, as producing the least ripple within the output channels of an auditory filter model (Holdsworth, et al., 1988). "Ripple" in this case was defined as the maximum ratio between absolute amplitudes averaged over two adjacent windows, one half period in length. It tends to be large if energy is concentrated in one half of the period. Fig. A-3 shows the ripple measure pattern across filter channels for Klatt, sine, and cosine phase, and for each of the 10 "random" phase patterns. The same "random" phase patterns were used throughout (in which sense they are not really random).

The pattern of partial frequencies of each inharmonic vowel was determined within the following constraints:

1) All partial frequencies are multiples of 4 Hz. This is to insure that 250 ms is a super-period of all beat patterns.

2) A partial must be at most F0/2, or 8*n Hz (where n is the partial's rank) whichever is smaller, from the harmonic series. This is to insure that the spectral density remains similar to that of a harmonic stimulus,

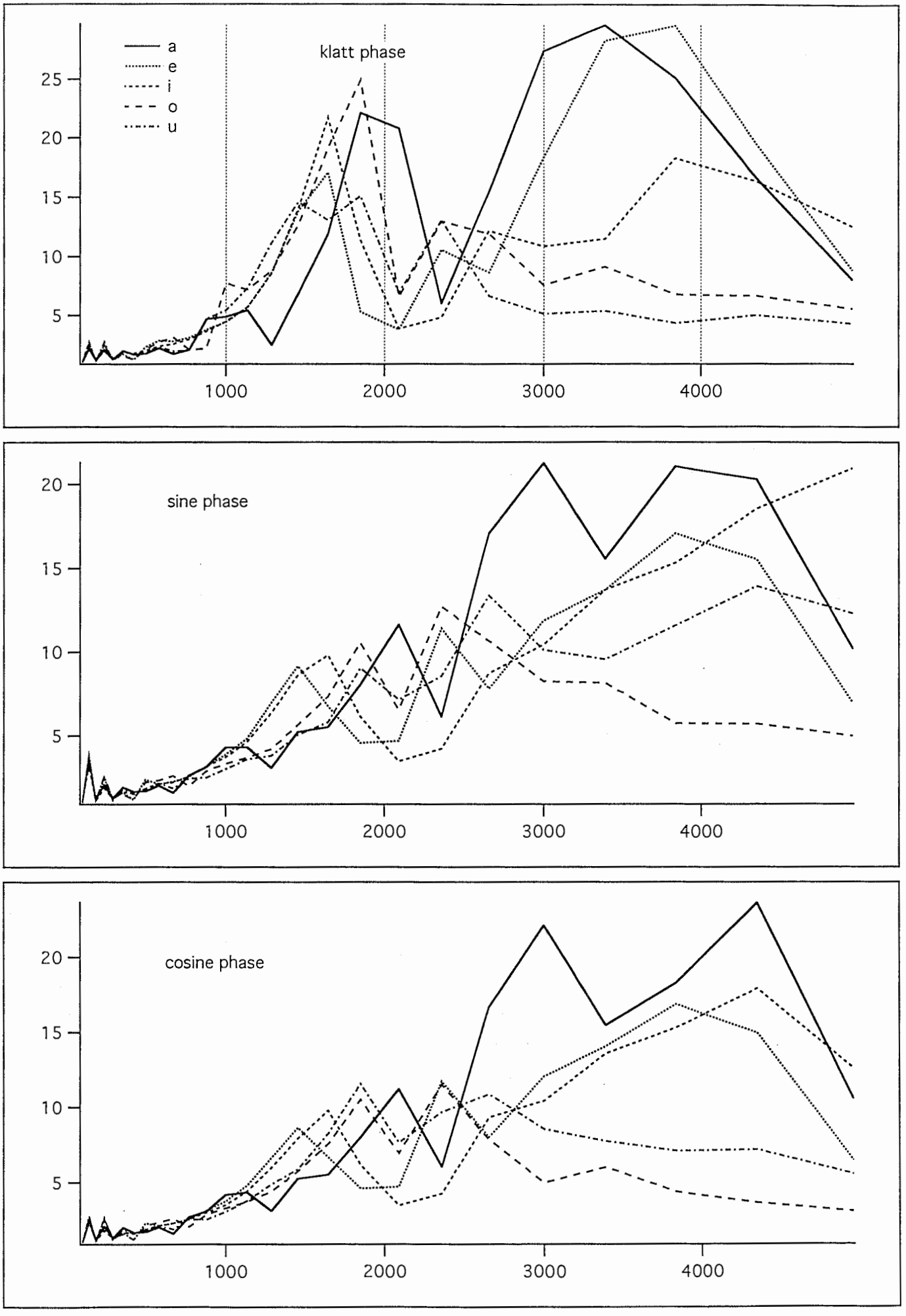
3) Each partial must be at least 16 Hz from:
   a) the frequency of the previous partial of the series,
   b) the frequency of any partial of the *other* vowel,
   c) the mirror images of these frequencies relative to the partial's harmonic frequency (to avoid any systematic shift)
4) Within these constraints, the partial is chosen randomly.

In order to satisfy constraint 3, different patterns were synthesized for both of the nominal frequencies used. When both vowels were inharmonic, their partials were randomly chosen by pairs that jointly satisfied the previous constraints, and assigned at random to either vowel. Constraint 3 was relaxed for the second harmonic (as it was incompatible with constraint 2 at that frequency).

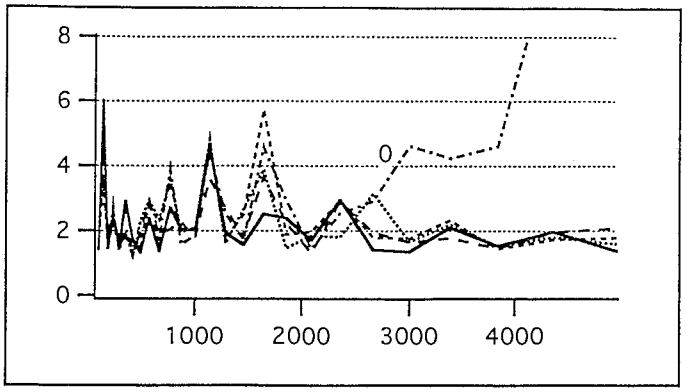
Finally, a measure of inharmonicity was defined as the sum of absolute differences between consecutive partial frequencies divided by their rank. This measure is sensitive to local rather than cross-spectrum harmonicity patterns, and puts relatively less weight on higher partials. For each condition 30 inharmonic patterns were produced and screened according to this measure, and the best chosen. The frequency of an inharmonic vowel is by convention the frequency of the harmonic series on which it is based.

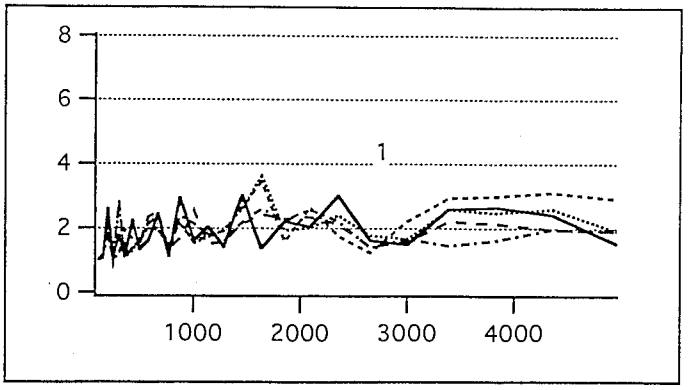
Fig. A-3 (next two pages): Ripple at the output of an auditory filter bank (Holdsworth et al., 1988), as a function of channel frequency, for all five vowels synthesized at 100 Hz, and for various phase patterns.

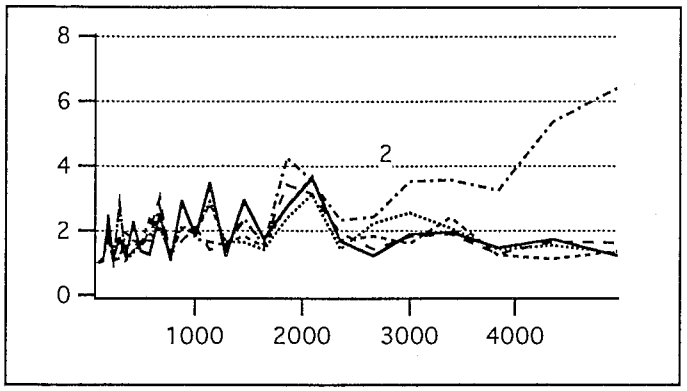
8 Mar 94 filter bank channel output modulation.

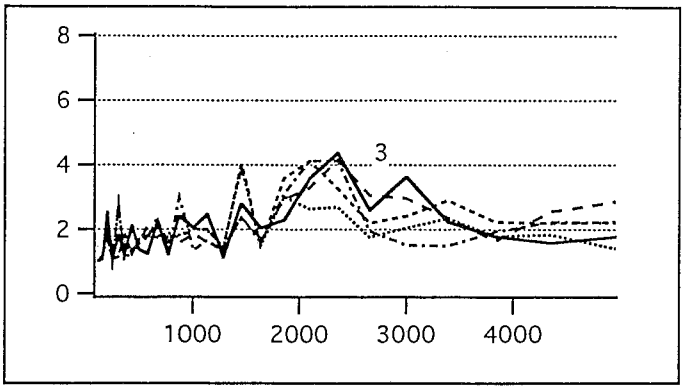


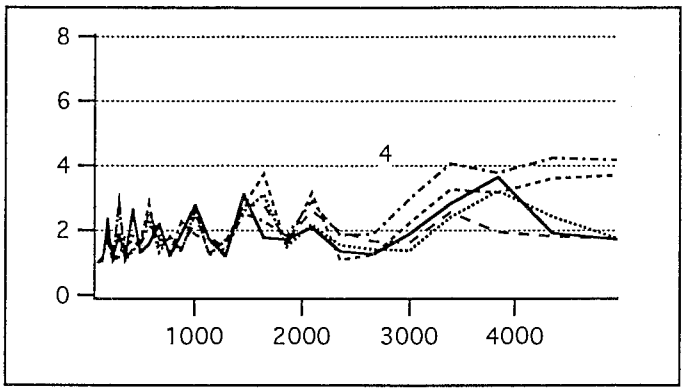
9 Mar 94 random phases, .125 Hz

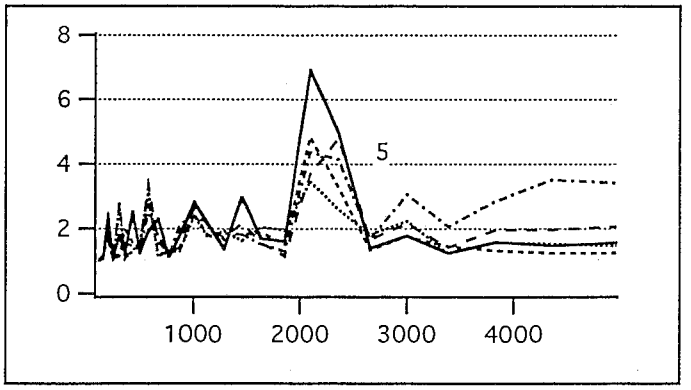


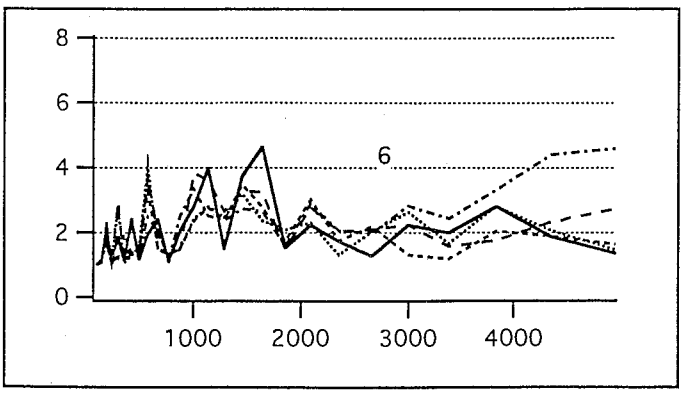


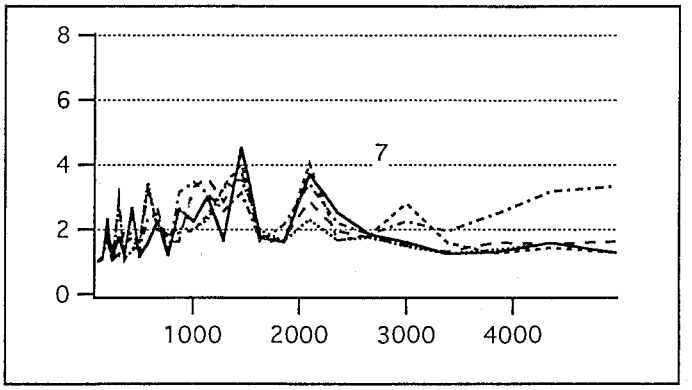


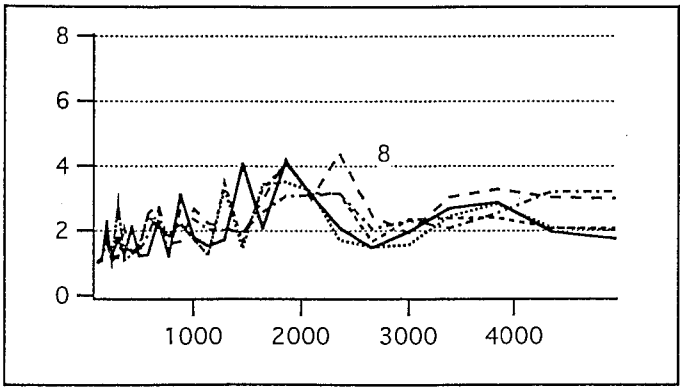


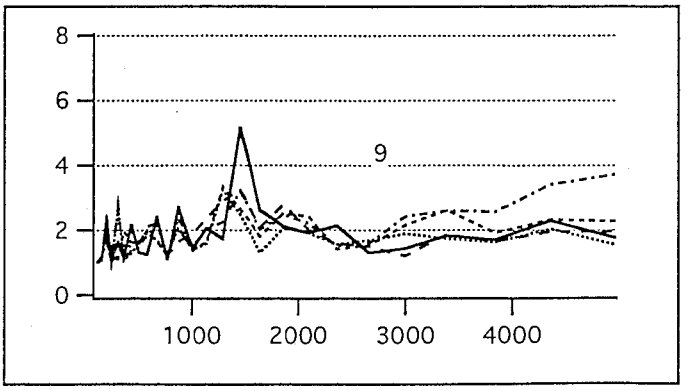












Appendix B. Vowel-specific effects in Experiment 1.

Fig. B-1 shows the identification rate and number of vowels responded for individual vowel pairs, averaged over subjects and sessions. There are considerable differences between pairs.

At unison, the plots for some pairs (/io/, /eo/, /eu/, /ae/, etc.) do not cross at 0 dB relative level, suggesting that within these pairs one vowel dominates the other. To explain such particularities, a first guess would be that our procedure of matching vowels by RMS. produced a loudness mismatch. However, insofar as loudness is reflected by A-weighted SPL measurements (Table A-1), a loudness mismatch within /ae/ for example should have produced a shift in the opposite direction.

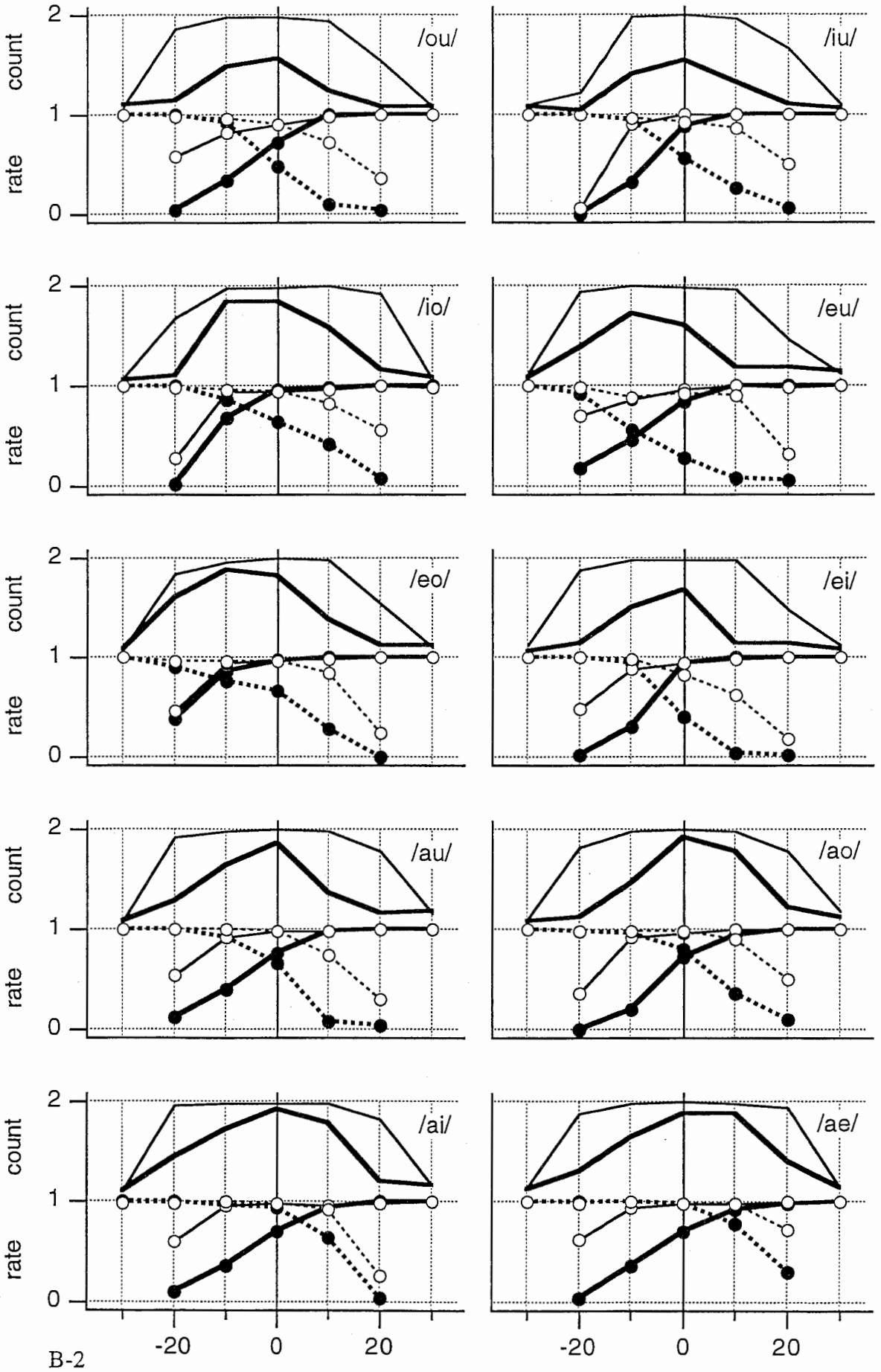
Another hypothesis is that identification of a vowel starts to degrade at a level such that the spectral peaks of formant F1 (resp. F2) merges with the other vowel's spectrum. To test this idea, we formed a new parameter by subtracting from the inter-vowel level parameter the particular value at which formant F1 (resp. F2) disappears into the spectral envelope of the ground. Graphically the parameter can be seen as the vertical separation between target and background spectral envelopes at F1 (resp. F2) of the target in Fig. A-1. If our conjecture is correct, these parameters should be good predictors of performance. To test the idea, we formed a linear model that fit identification by either level, or our two new parameters (together with ΔF_0). Despite a larger number of parameters, the two new parameters produce a less good fit than simply level (r^2 = 0.45 vs. 0.51). Evidently this simple "formant disappearing" model is inadequate.

The effect of ΔF_0 on identification varies from vowel to vowel. For example it is large for /ue/ (descending lines in graph labeled /eu/), and small for /eo/. Similar differences appear for the number of vowels responded. The asymmetry in effects between vowels in a pair contradicts the principle of symmetric segregation that is assumed to hold for primary segregation mechanisms (Bregman, 1990).

To attempt to explain vowel-pair specificities in ΔF_0 effect, let us assume that segregation occurs according to the beats hypothesis: identification improves when there are strong beats near important formants. Such strong beats should occur when target and background envelopes have similar levels. To test this hypothesis we again formed two new parameters for each target-ground pair, this time representing the *absolute* difference between envelopes at formant F1 (resp. F2) of the target. We compared these parameters to a single parameter formed by subtracting -10 dB from the level parameter and taking the absolute value. Again, the fit was less good despite the larger number of parameters (r^2=0.20 vs. 0.26). This model is clearly inadequate, possibly because it is too crude, and possibly because beats do not determine segregation in this case.

Meddis's model of concurrent vowel identification sorts peripheral auditory channels according to whether or not they respond with the periodicity of the dominant vowel (the one whose periodicity dominates the overall response) (Meddis and Hewitt., 1992). If all channels are dominated by the same periodicity, the model cannot work, so ΔF_0 should have no effect. This might be the case if there is a large level mismatch {this modeling remains to be done...}

Fig. B-1 (next page). Identification rate (lines with markers), and number of vowels responded (lines) as a function of inter-vowel level, for F_0 differences of 0% (thick lines and filled symbols) and 6% (thin lines and open symbols) and for all vowel pairs. The identification rate of the first vowel of each pair is represented by the continuous (ascending) lines, that of the second by the dotted (descending) lines. Extreme points (∞) correspond to single vowel conditions.


/ou/
/iu/
/io/
/eu/
/eo/
/ei/
/au/
/ao/
/ai/
/ae/
count
rate
2
1
0
-20
0
20

Appendix C Subject-specific effects in Experiment 1.

Fig. C-1. shows the number of vowels responded and identification rate for all subjects. The most visible difference is the degree to which ΔF_0 affects identification or the number of vowels responded. Effects are large for subject T, and small for subject K. The smaller effects for K are partly (but not entirely) due to higher scores at unison. An explanation is that this subject ignored the multiplicity cue and systematically gave two responses, as suggested by the relatively high number of vowels responded for single vowels.

Identification is evidently affected by the number of vowels responded, and thus indirectly by "multiplicity cues". Quite interesting in this respect are identification rates conditional on two responses (Fig. C-2), and the results of Experiment 5 in which multiplicity cues were thwarted by the use of the two-response task.

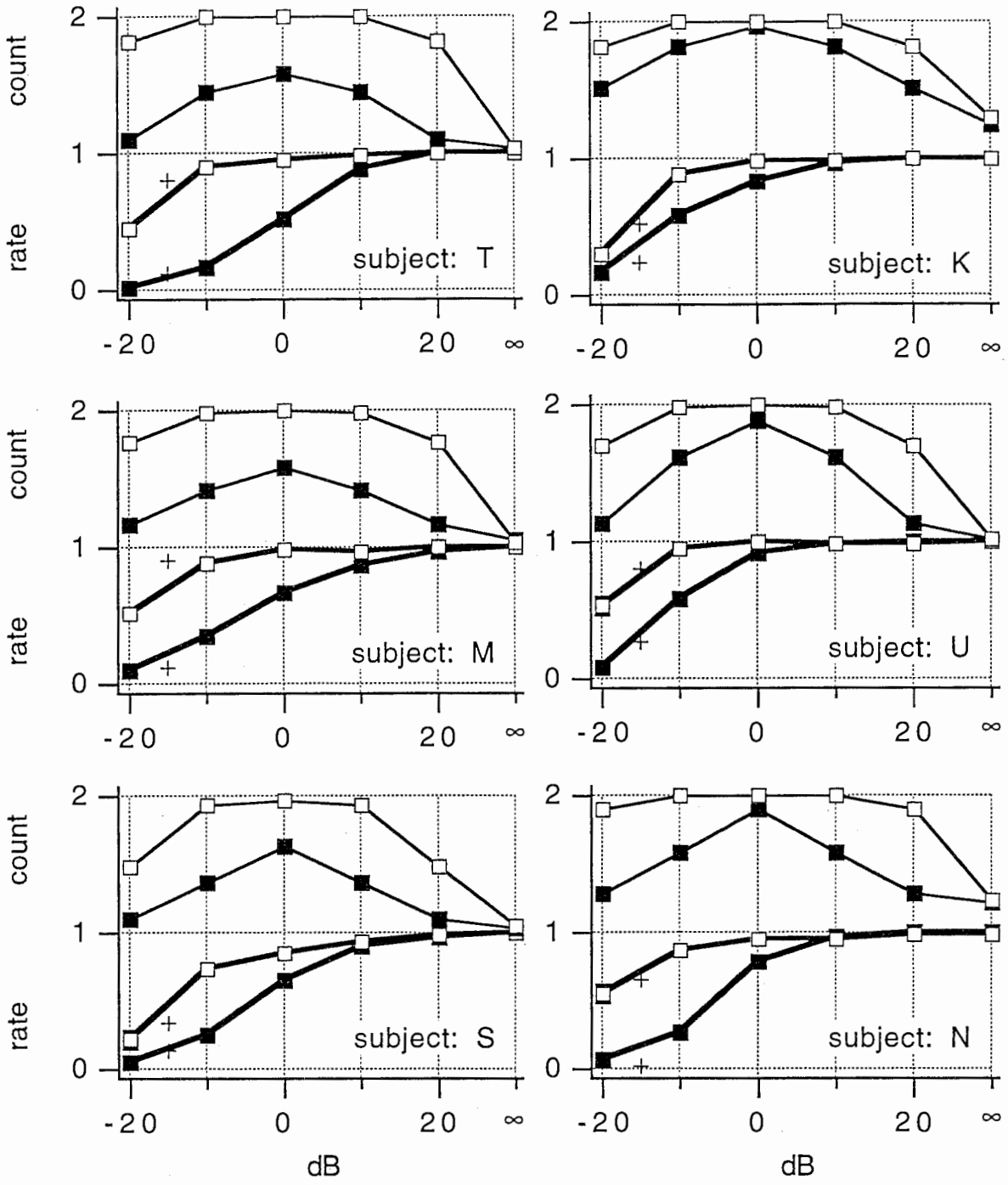


Fig. C-1. Number of vowels responded (thin lines) and identification rates (thick lines) as a function of relative level, at unison (filled markers) and at a ΔF_0 of 6% (open markers), for each subject. Crosses are results obtained in Exp. 2 (target at -15 dB) for ΔF_0s of 0 and 6.45%.

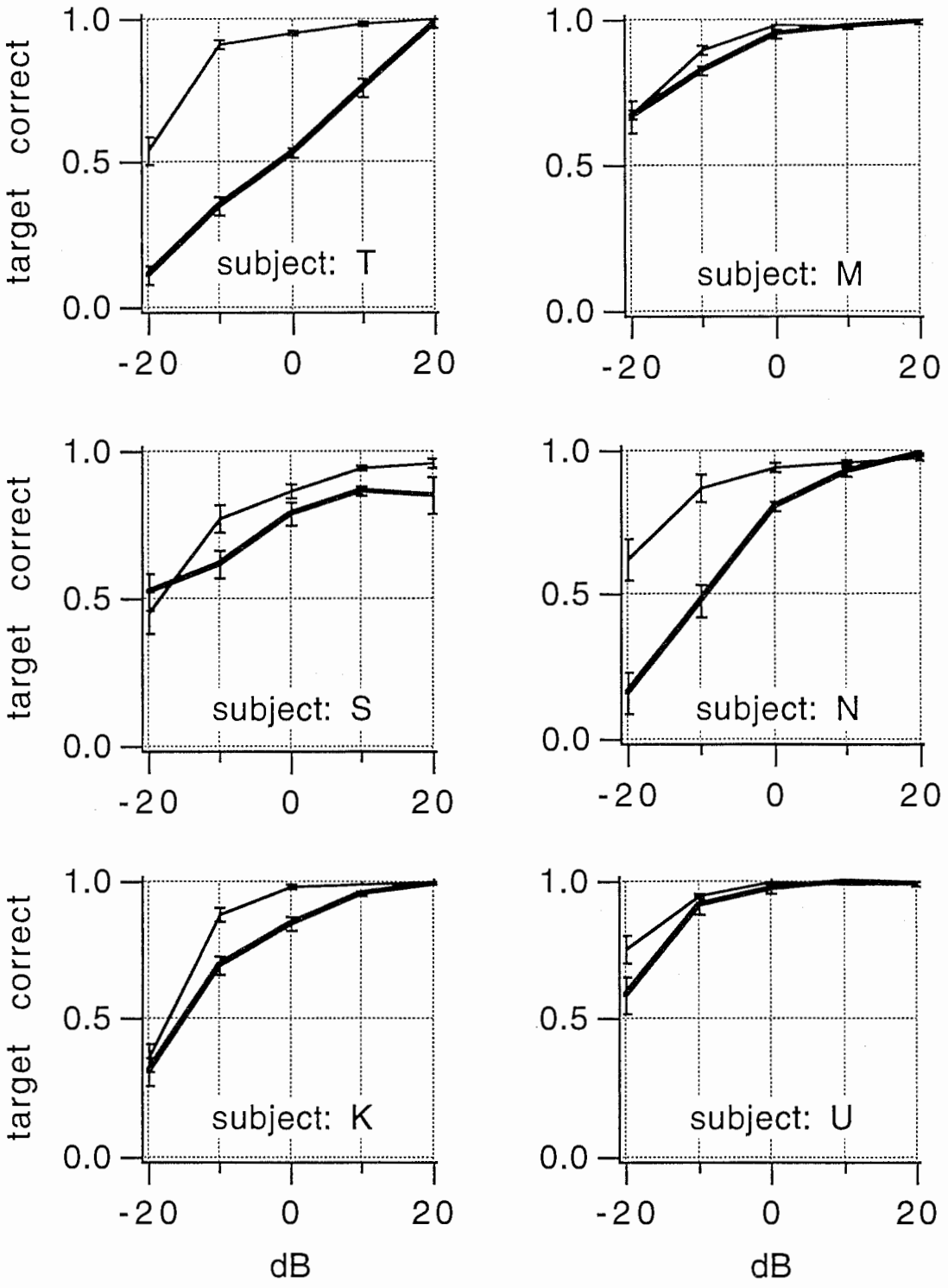
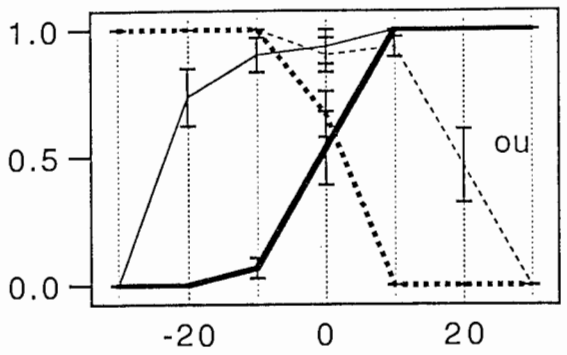
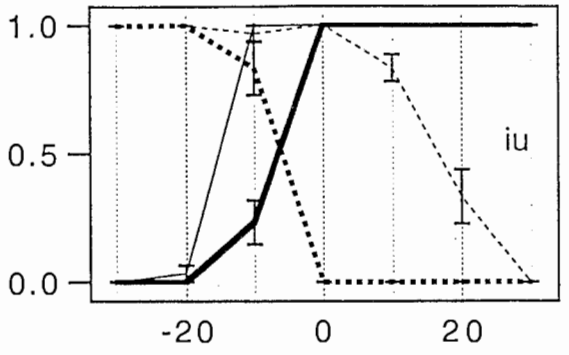
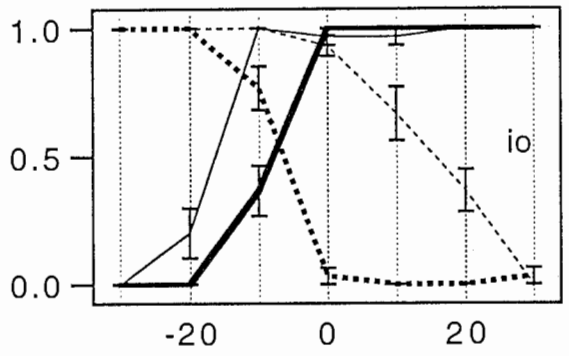


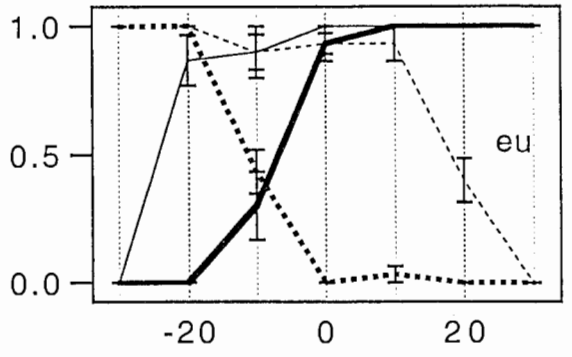
Fig. C-2. Identification rates conditional on a two-vowel response, as a function of relative level between vowels, for all six subjects. Thick lines are for unison, thin lines are for 6.45% ΔF_0.

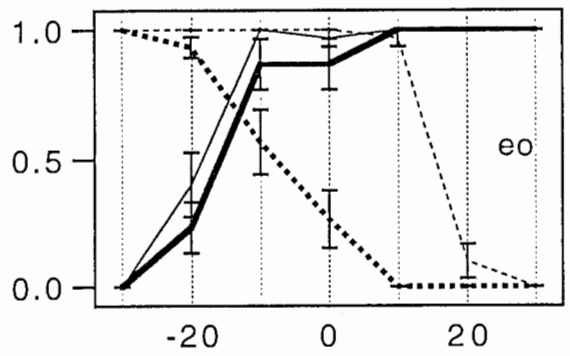
Fig. C-3 (next pages): Identification rate as a function of relative level at unison (thick lines) and with a ΔF_0 of 6.45% (thin lines), for each vowel pair and each individual subject. Extreme points with values near 1.0 represent the target vowel alone. Extreme points with values near 0 are meaningless. Error bars represent standard error over sessions.

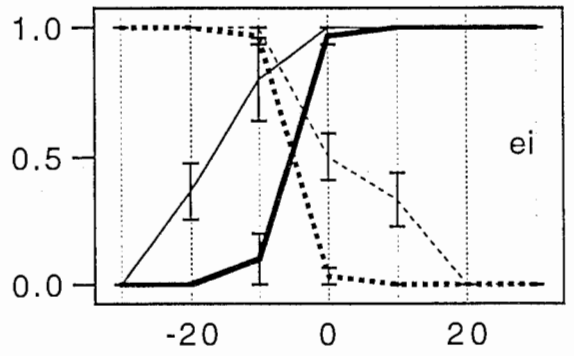

1.0
0.5
0.0
-20
0
20
ou

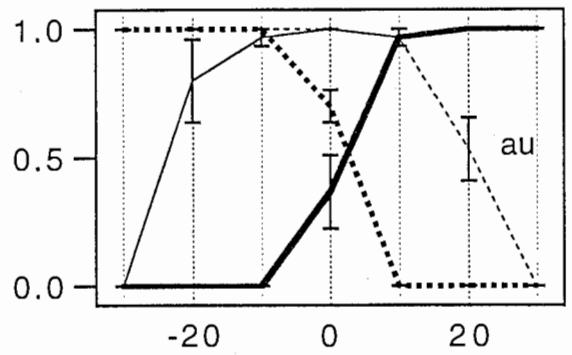

1.0
0.5
0.0
-20
0
20
iu

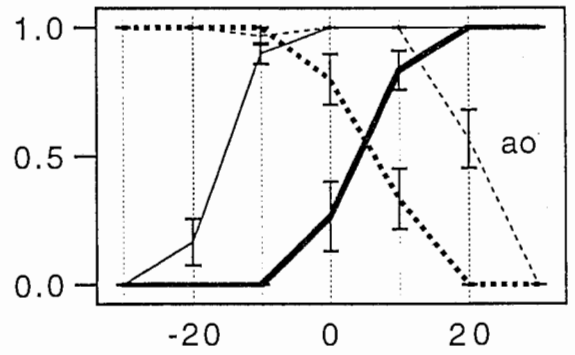

1.0
0.5
0.0
-20
0
20
io

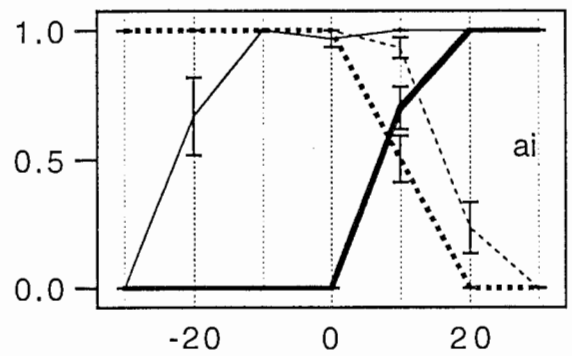

1.0
0.5
0.0
-20
0
20
eu

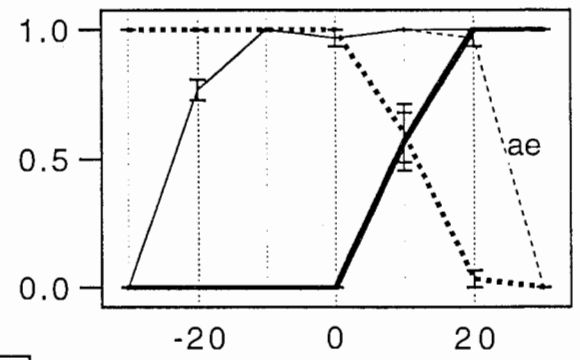

1.0
0.5
0.0
-20
0
20
eo


1.0
0.5
0.0
-20
0
20
ei

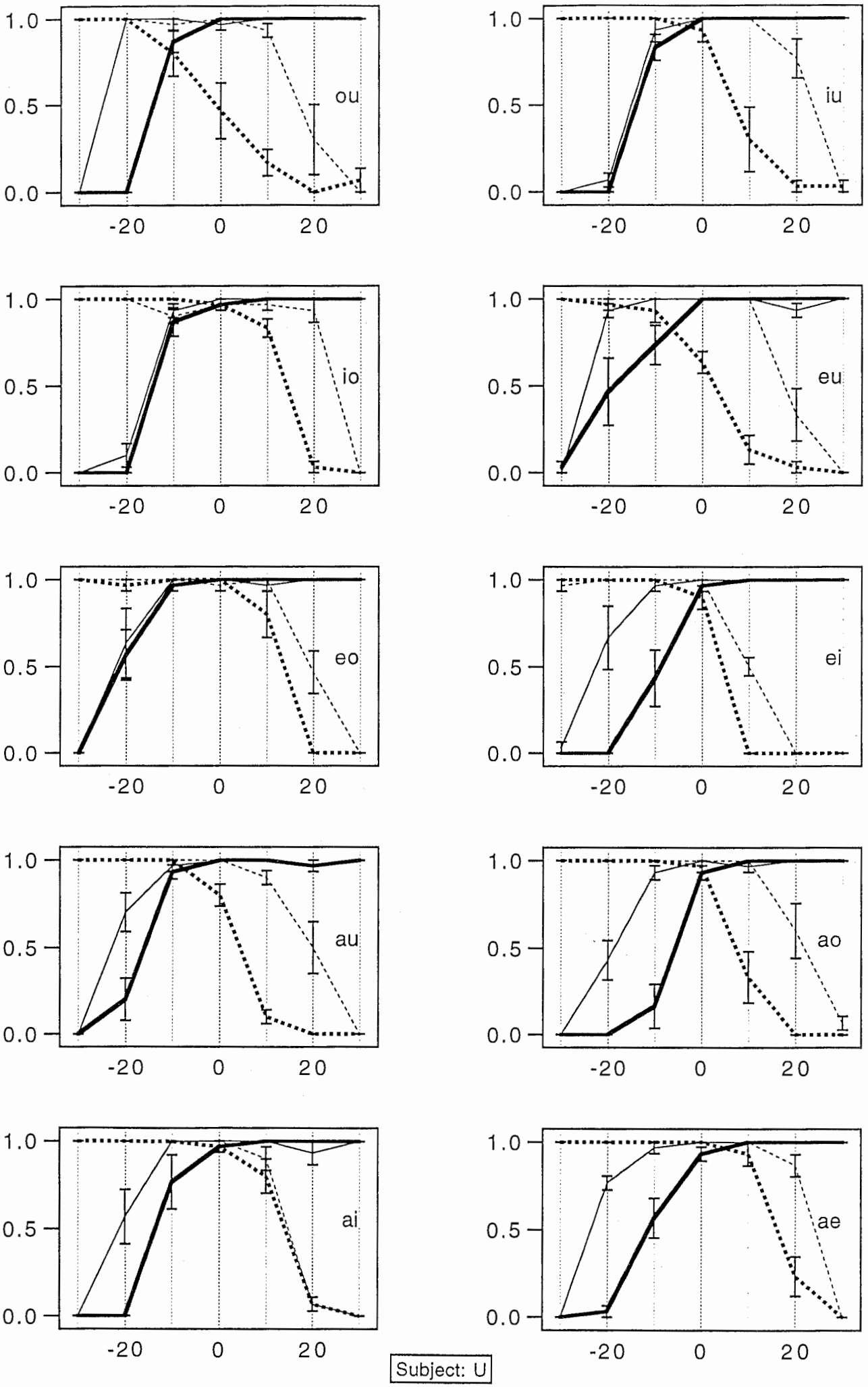

1.0
0.5
0.0
-20
0
20
au

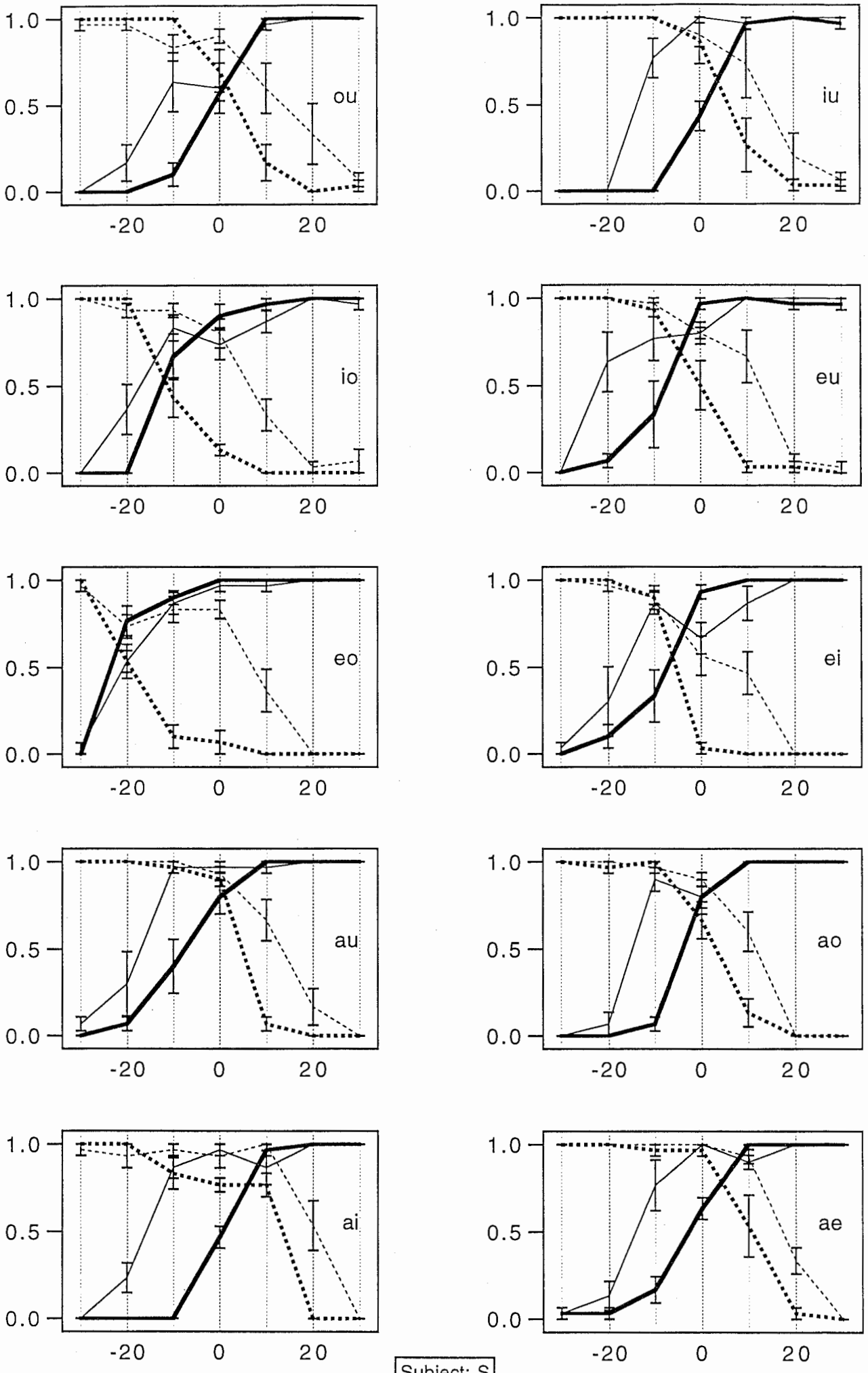

1.0
0.5
0.0
-20
0
20
ao


1.0
0.5
0.0
-20
0
20
ai

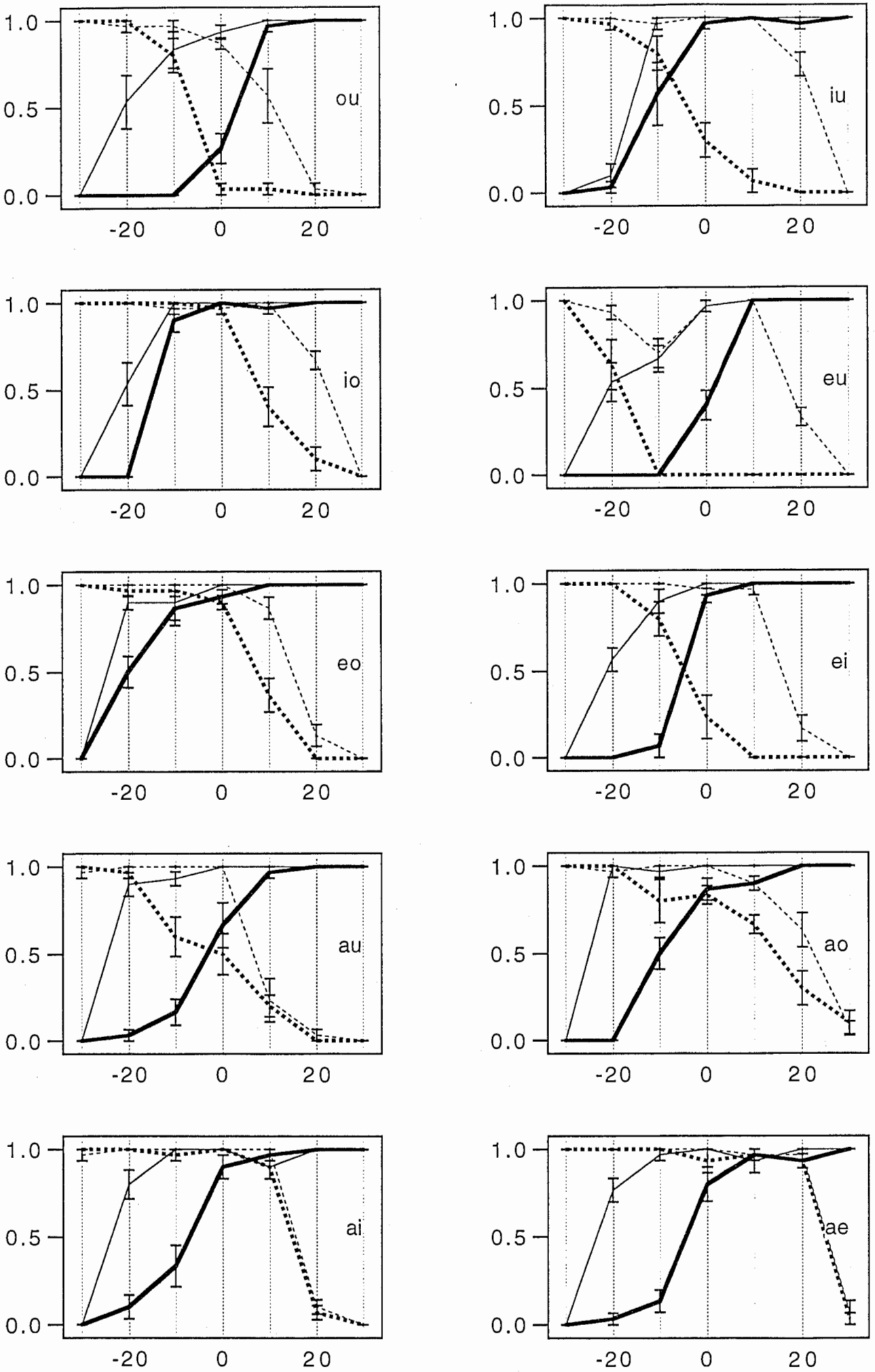

1.0
0.5
0.0
-20
0
20
ae

Subject: T

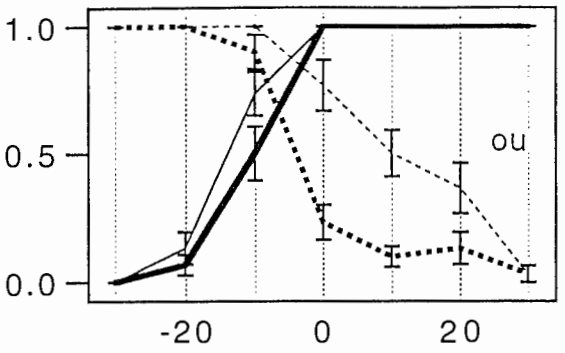

1.0
0.5
0.0
-20
0
20
ou
iu
io
eu
eo
ei
au
ao
ai
ae
Subject: U

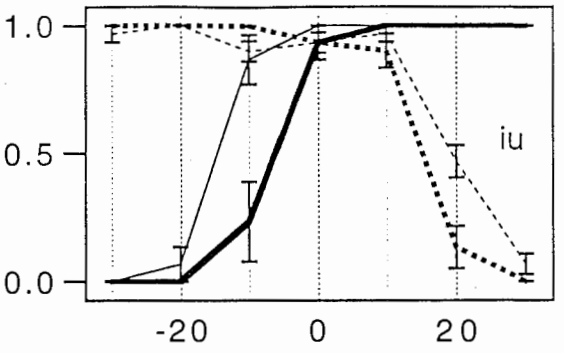


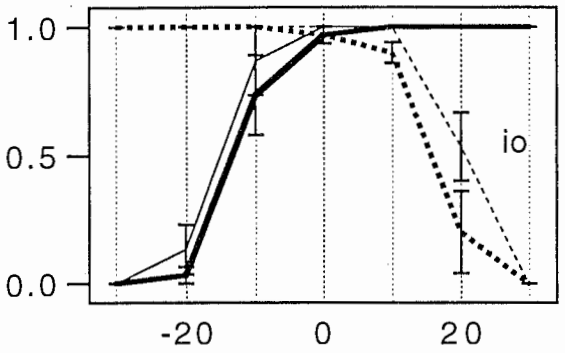
Subject: S

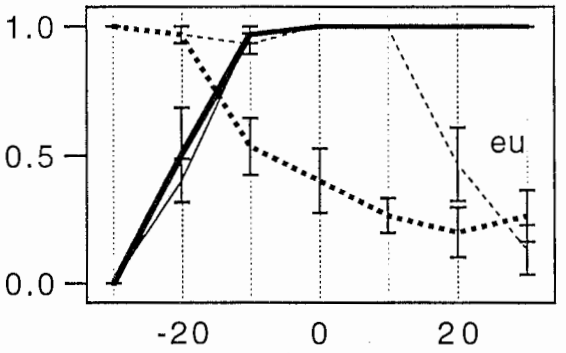


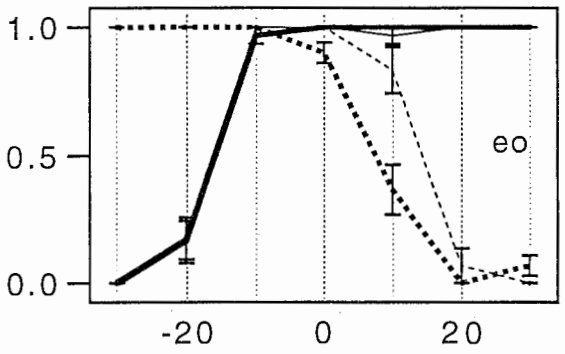
Subject: M

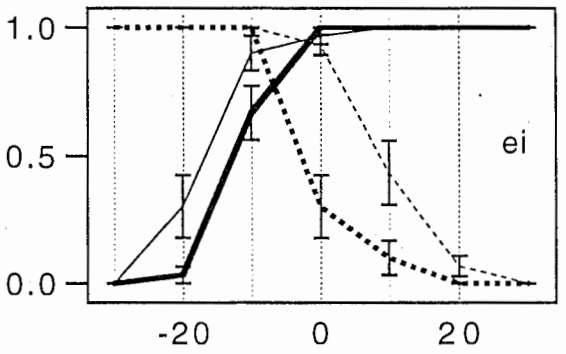

1.0
0.5
0.0
-20
0
20
ou

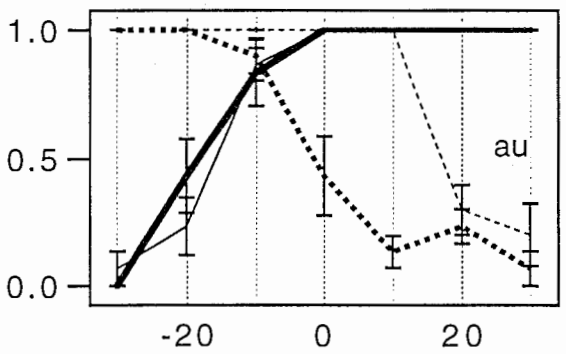

1.0
0.5
0.0
-20
0
20
iu

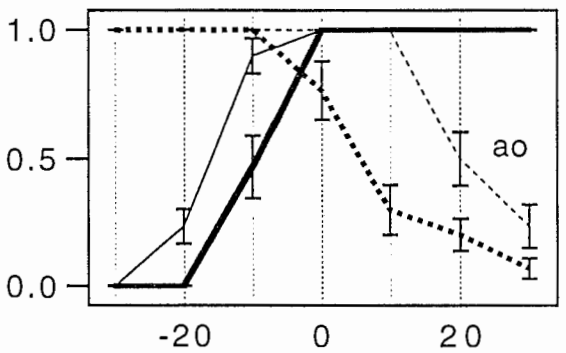

1.0
0.5
0.0
-20
0
20
io

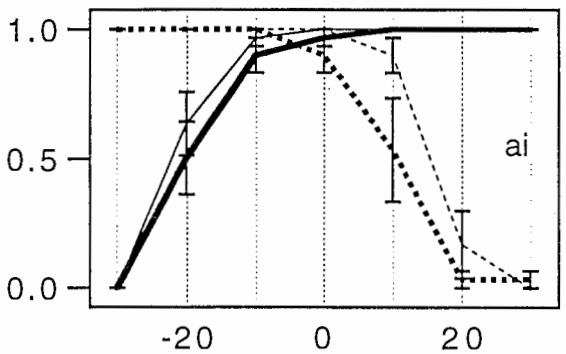

1.0
0.5
0.0
-20
0
20
eu

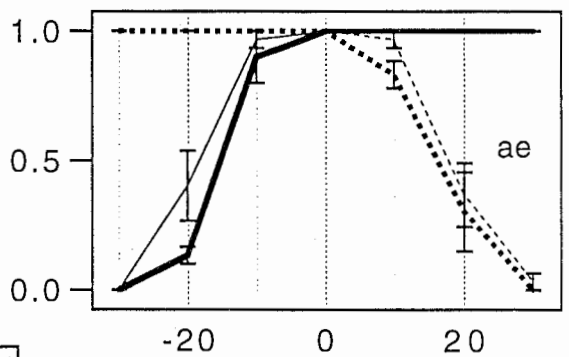

1.0
0.5
0.0
-20
0
20
eo


1.0
0.5
0.0
-20
0
20
ei


1.0
0.5
0.0
-20
0
20
au


1.0
0.5
0.0
-20
0
20
ao


1.0
0.5
0.0
-20
0
20
ai


1.0
0.5
0.0
-20
0
20
ae

Subject: K

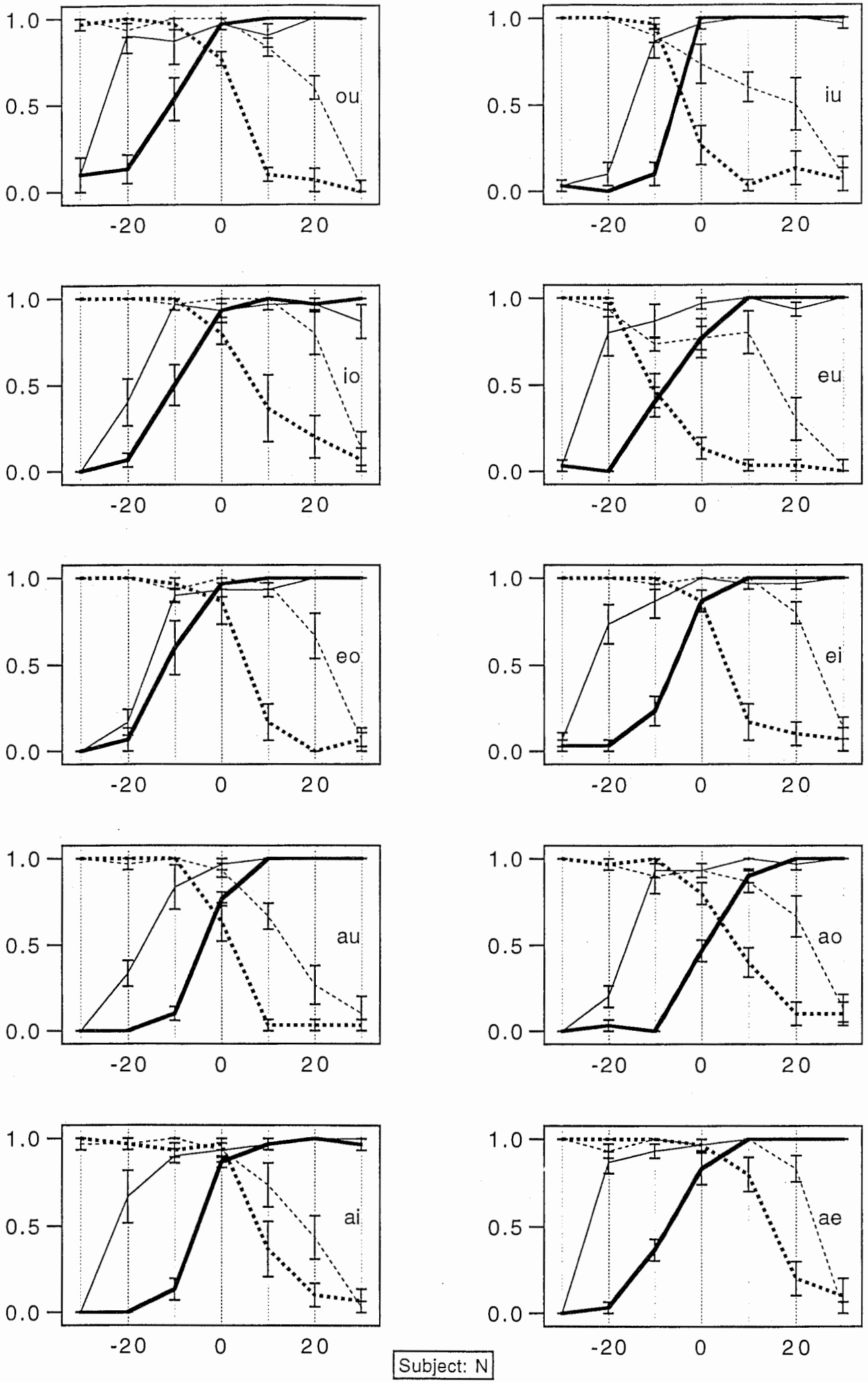
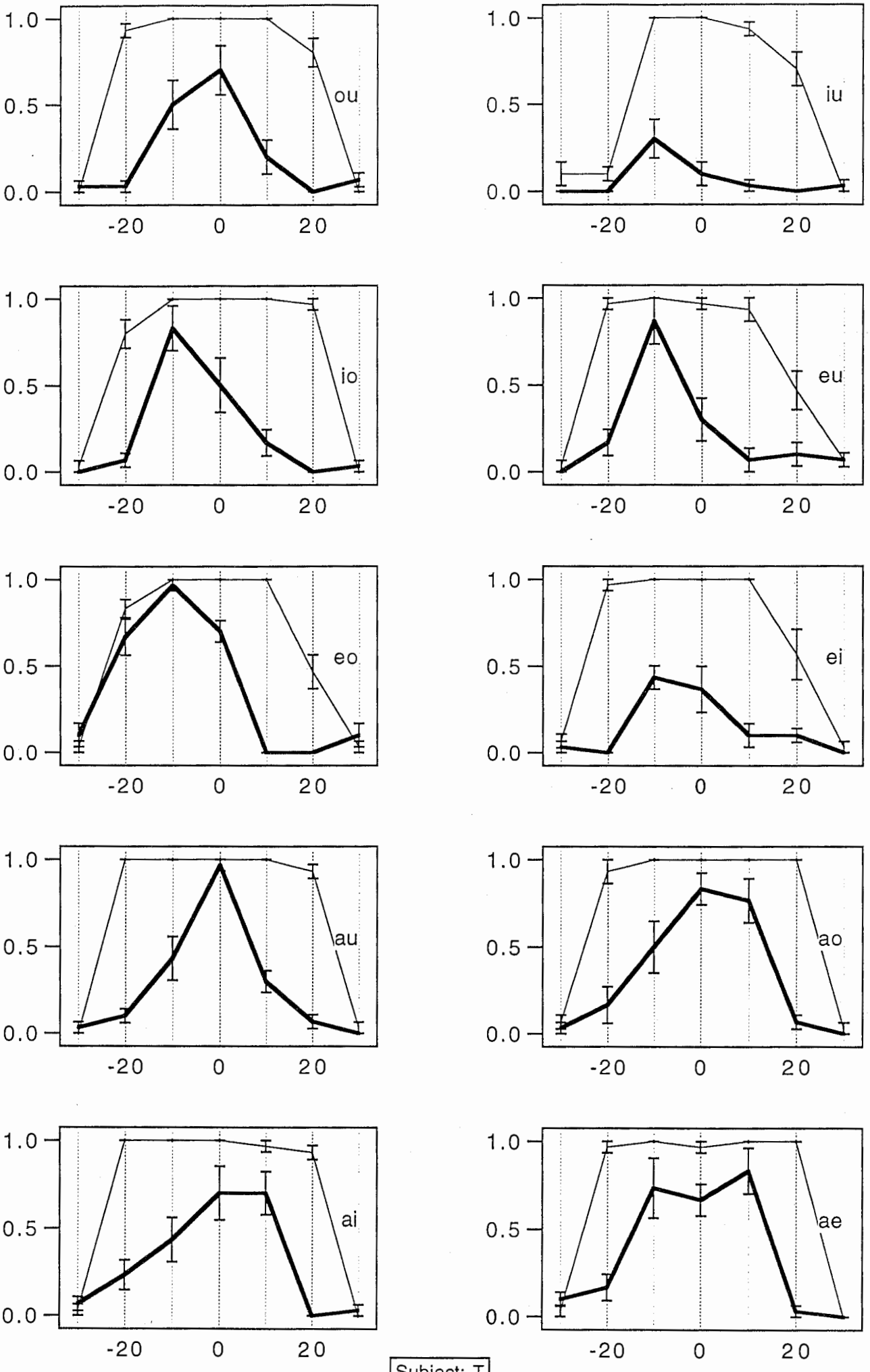
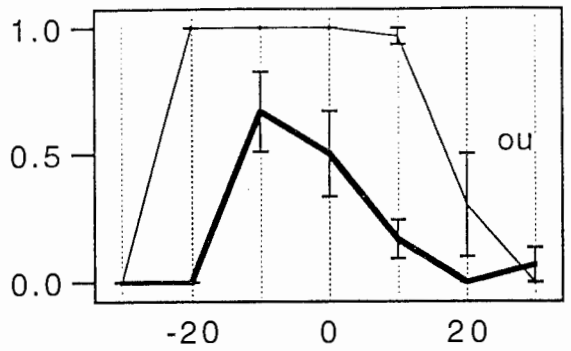
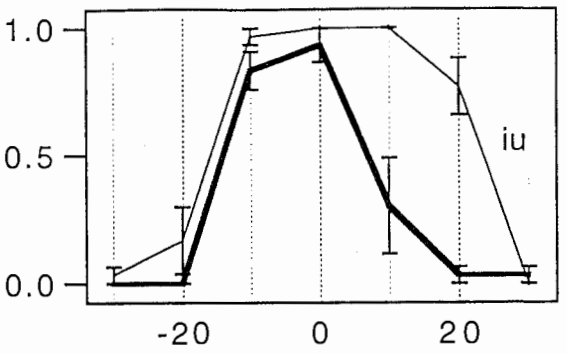
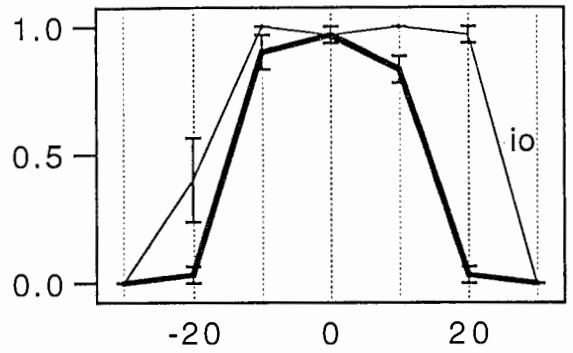

1.0
0.5
0.0
-20
0
20
ou
iu
io
eu
eo
ei
au
ao
ai
ae
Subject: N

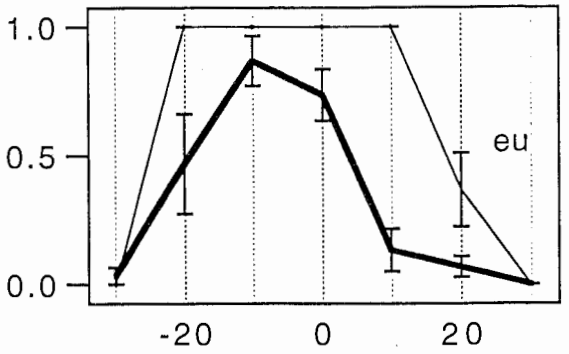
Fig C-4 (next pages): Number of vowels responded as a function of relative level at unison (thick lines) and with a ΔF_0 of 6.45% (thin lines), for each vowel pair and each individual subject. Extreme points represent single vowels.

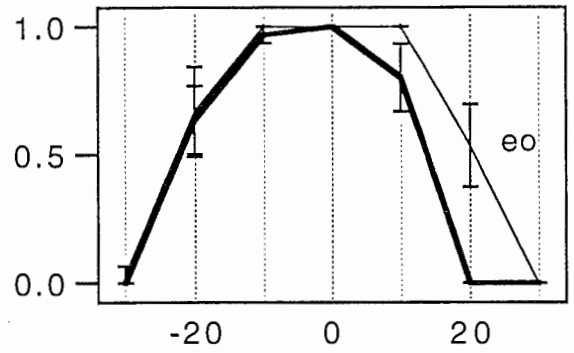

1.0
0.5
0.0
-20
0
20
ou
iu
io
eu
eo
ei
au
ao
ai
ae
Subject: T

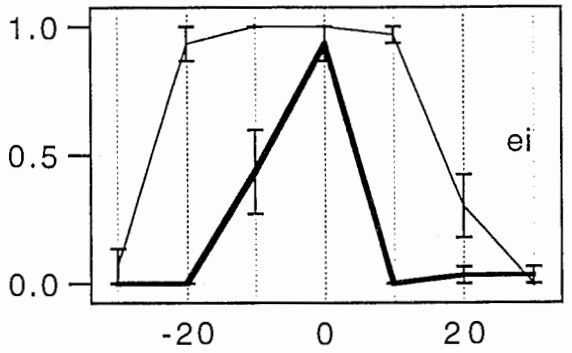

1.0
0.5
0.0
ou
-20
0
20

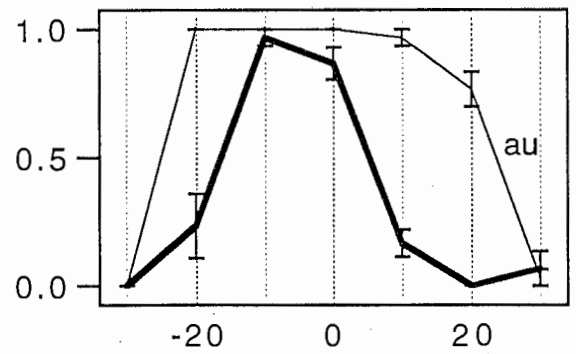

1.0
0.5
0.0
iu
-20
0
20

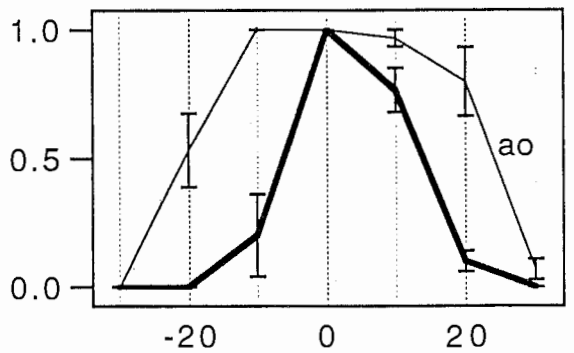

1.0
0.5
0.0
io
-20
0
20

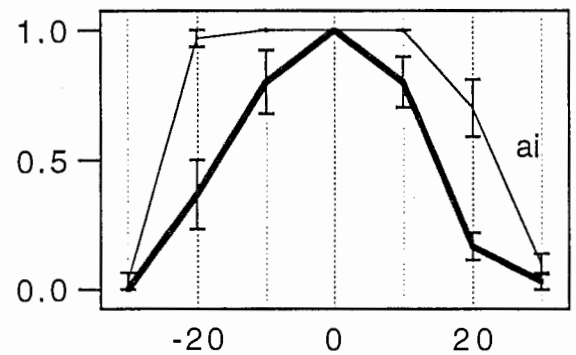

1.0
0.5
0.0
eu
-20
0
20

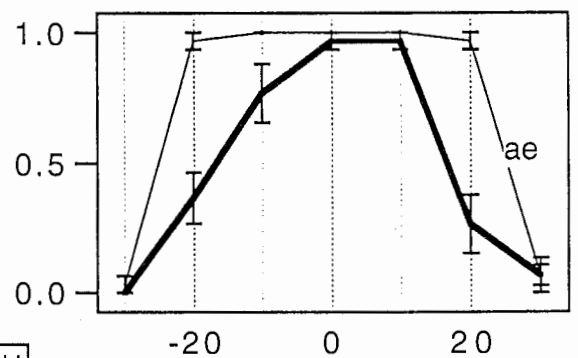

1.0
0.5
0.0
eo
-20
0
20


1.0
0.5
0.0
ei
-20
0
20

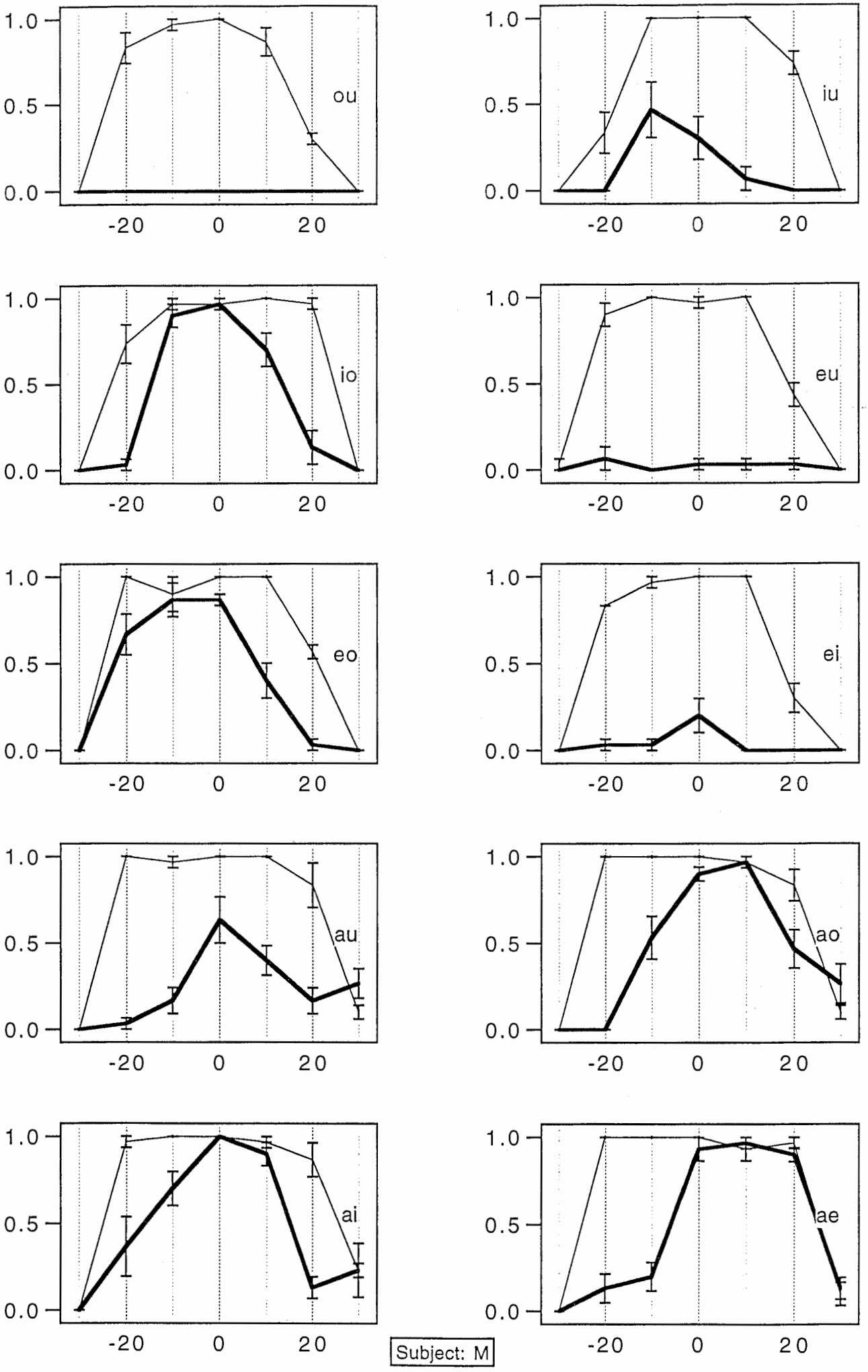

1.0
0.5
0.0
au
-20
0
20

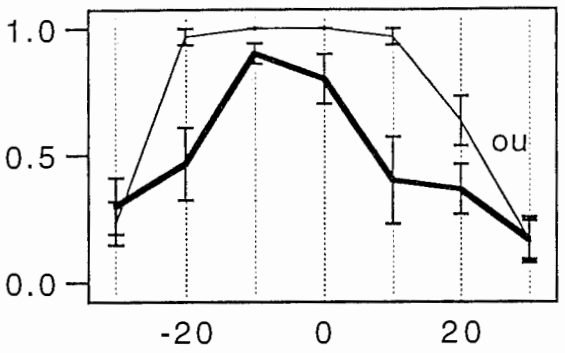

1.0
0.5
0.0
ao
-20
0
20

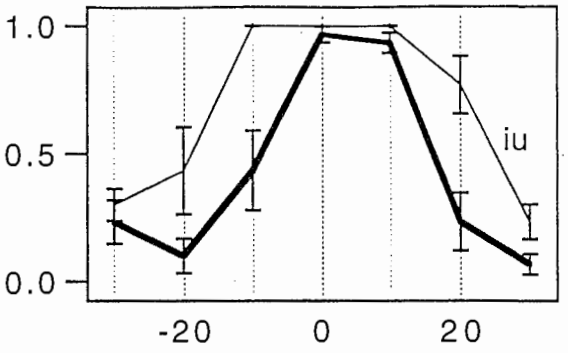

1.0
0.5
0.0
ai
-20
0
20

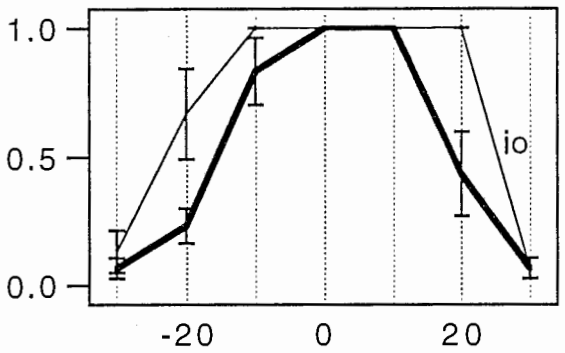

1.0
0.5
0.0
ae
-20
0
20

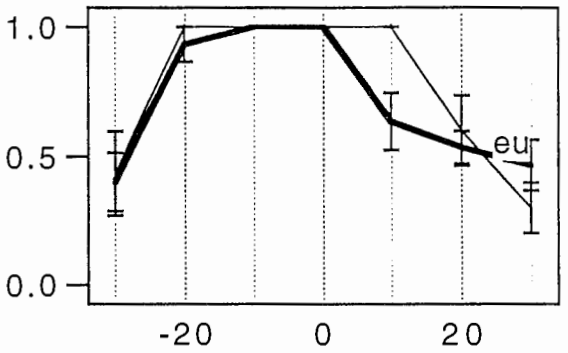
Subject: U

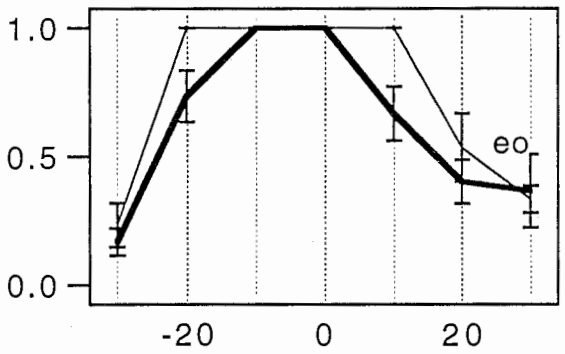

1.0
0.5
0.0
-20
0
20
ou
iu
io
eu
eo
ei
au
ao
ai
ae
Subject: M

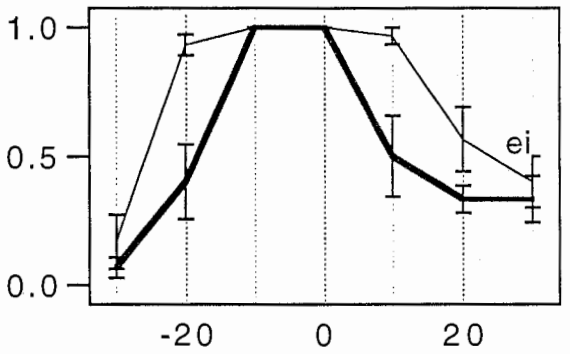

1.0
0.5
0.0
ou
-20
0
20

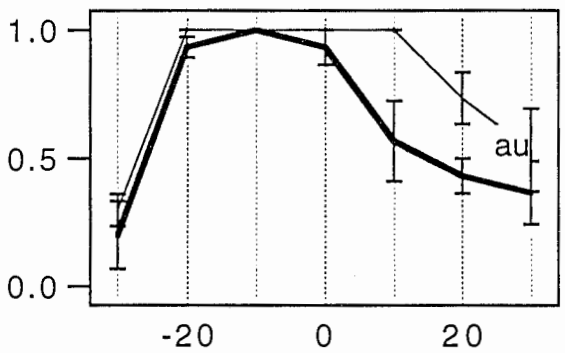

1.0
0.5
0.0
iu
-20
0
20

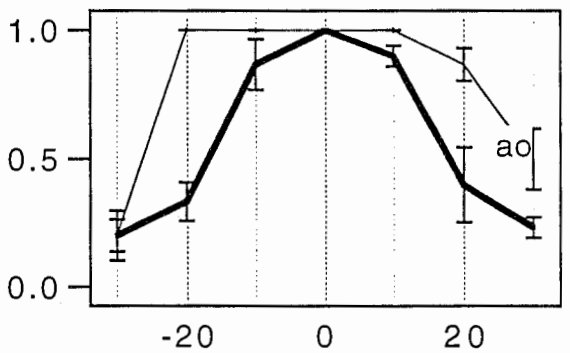

1.0
0.5
0.0
io
-20
0
20

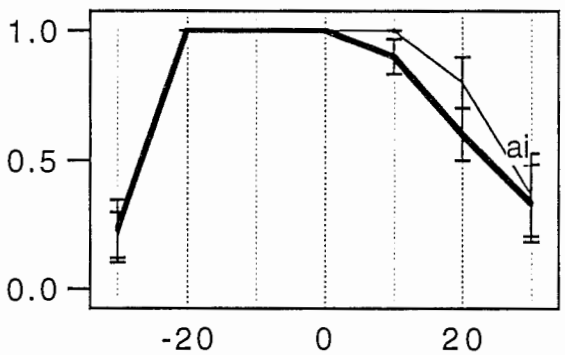

1.0
0.5
0.0
eu
-20
0
20

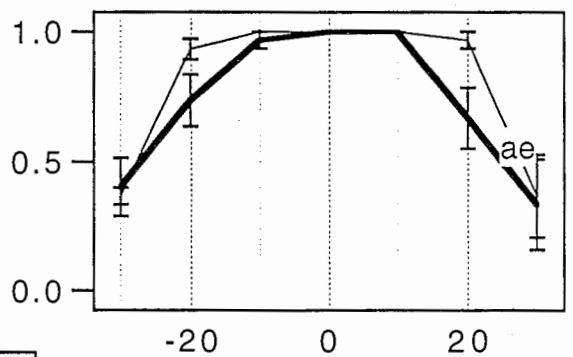

1.0
0.5
0.0
eo
-20
0
20


1.0
0.5
0.0
ei
-20
0
20

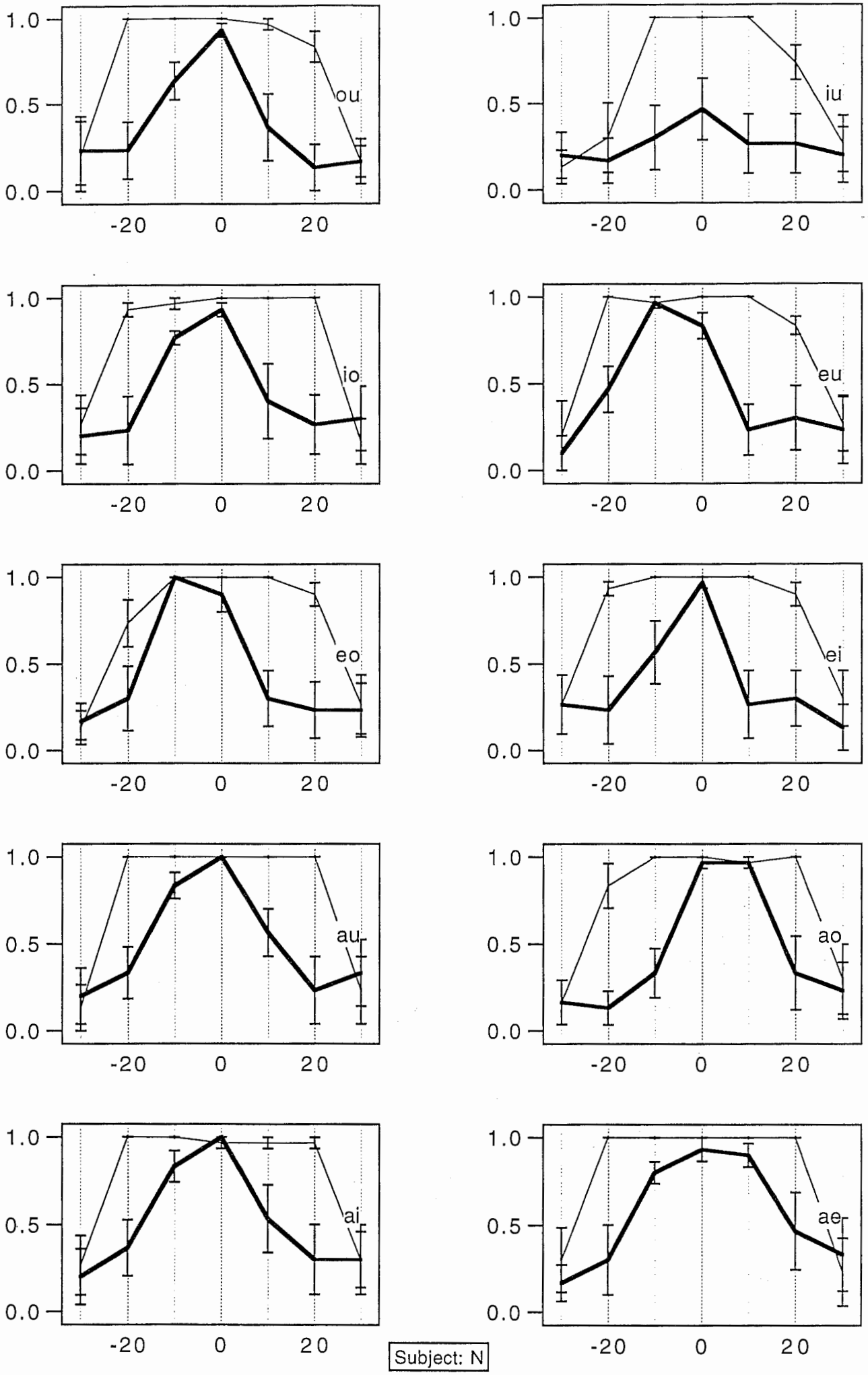

1.0
0.5
0.0
au
-20
0
20

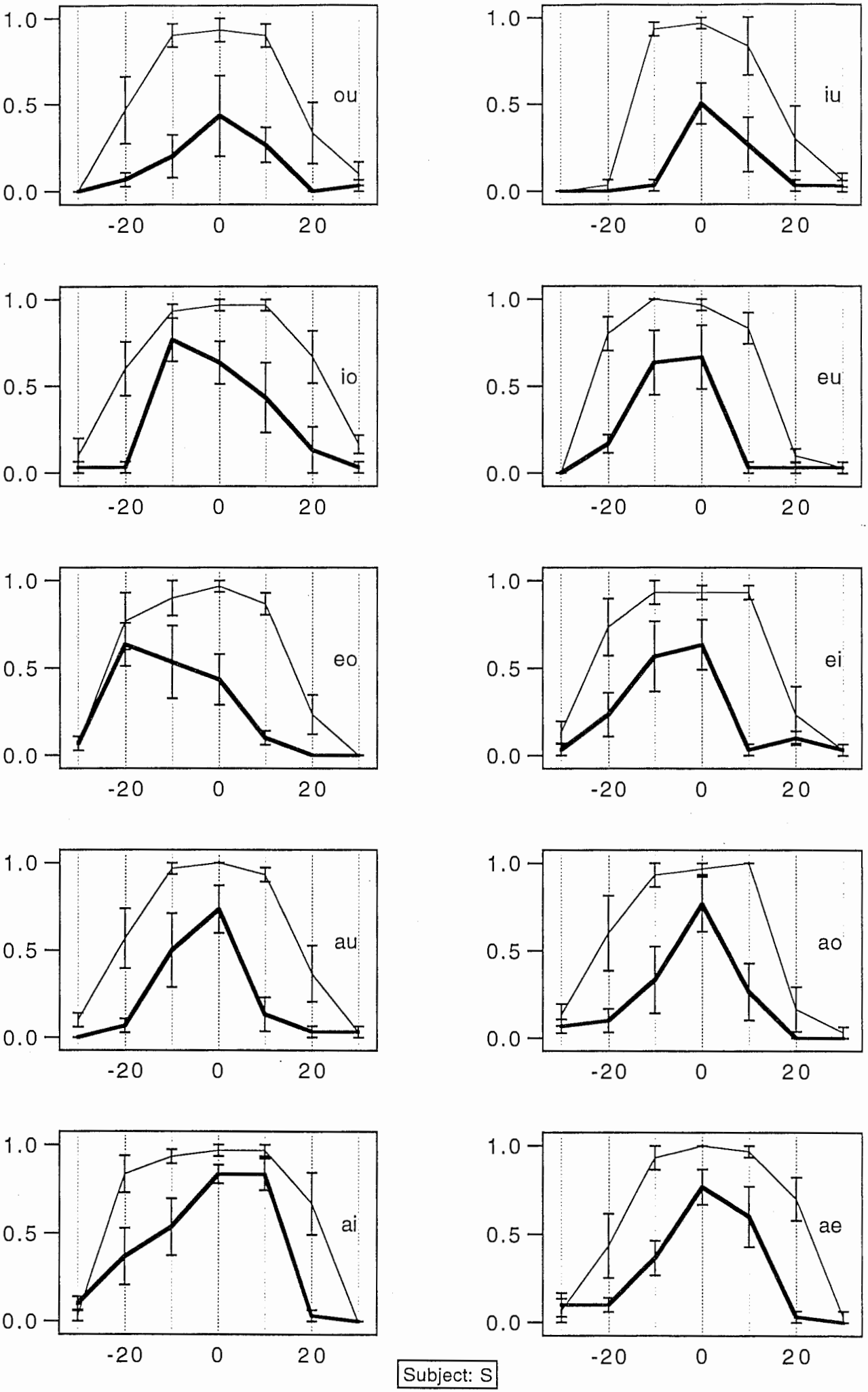

1.0
0.5
0.0
ao
-20
0
20


1.0
0.5
0.0
ai
-20
0
20


1.0
0.5
0.0
ae
-20
0
20

Subject: K


1.0
0.5
0.0
-20
0
20
ou
iu
io
eu
eo
ei
au
ao
ai
ae
Subject: N


1.0
0.5
0.0
-20
0
20
ou
iu
io
eu
eo
ei
au
ao
ai
ae
Subject: S

Appendix D. Session effects

Session-to-session variations were examined for any particularities related to the two different tasks used.

Experiment 1

Fig D-1 (left) shows the average identification rate over sessions. There is a significant improvement with session number ($p<.0001$), as indicated by a repeated measures analysis with fixed factors LEVEL, DELTA and SESSION (treated as a regressor), and random factor SUBJECT. No interaction was significant. Mean identification over the last two sessions is about 5% higher than over the first three.

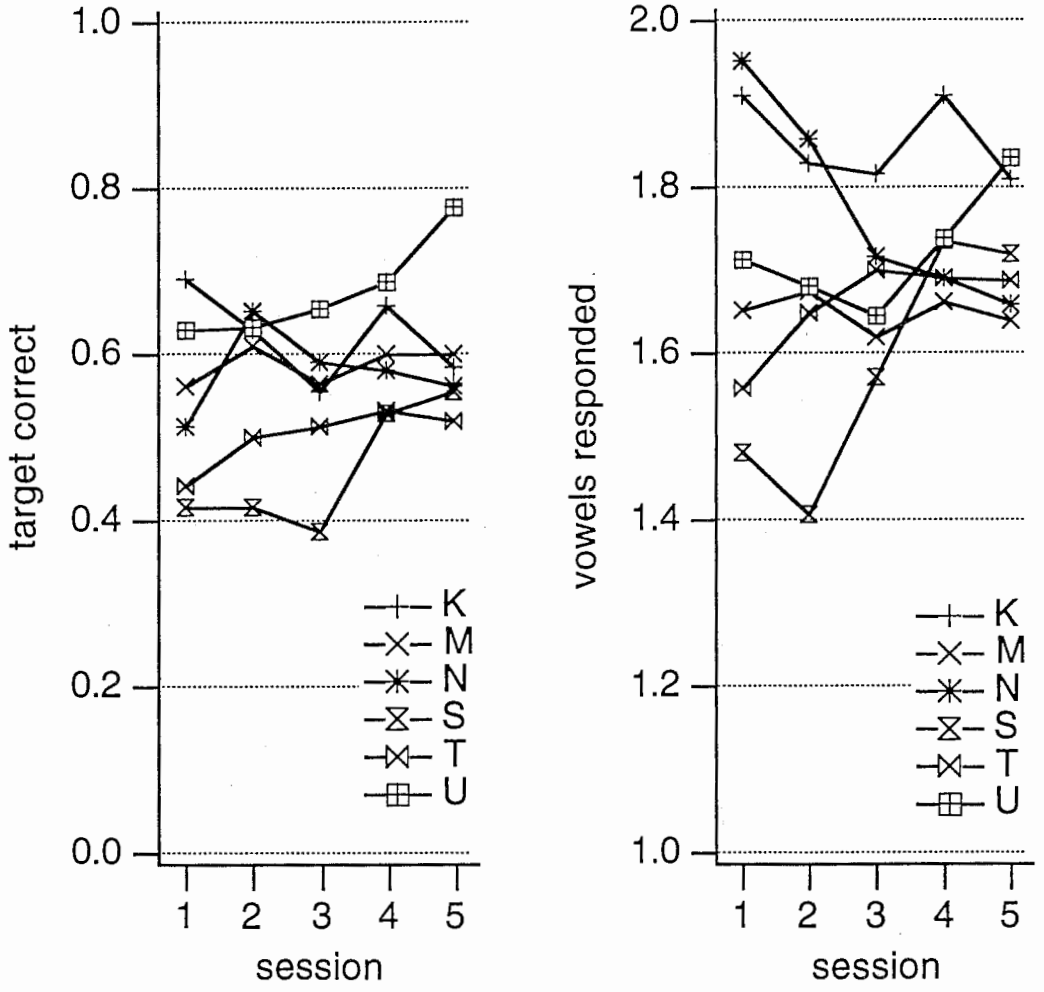


Fig. D-1. Identification rate (left) and number of vowels responded (right) as a function of session number in Exp. 1, for all subjects.

Experiments 2-4

Sessions were common to Experiments 2-4. Variations over sessions are illustrated in Fig D-2. The downward trend in the number of vowels responded is significant at the population level ($p<.0001$), as indicated by a repeated measures ANOVA with fixed factors CONDITION (ΔF_0 and harmonicity combined) and SESSION (treated as a regressor) and random factor SUBJECT. Interaction with SUBJECT was also significant ($p<.0001$). A similar analysis of the identification rate showed no significant trend.

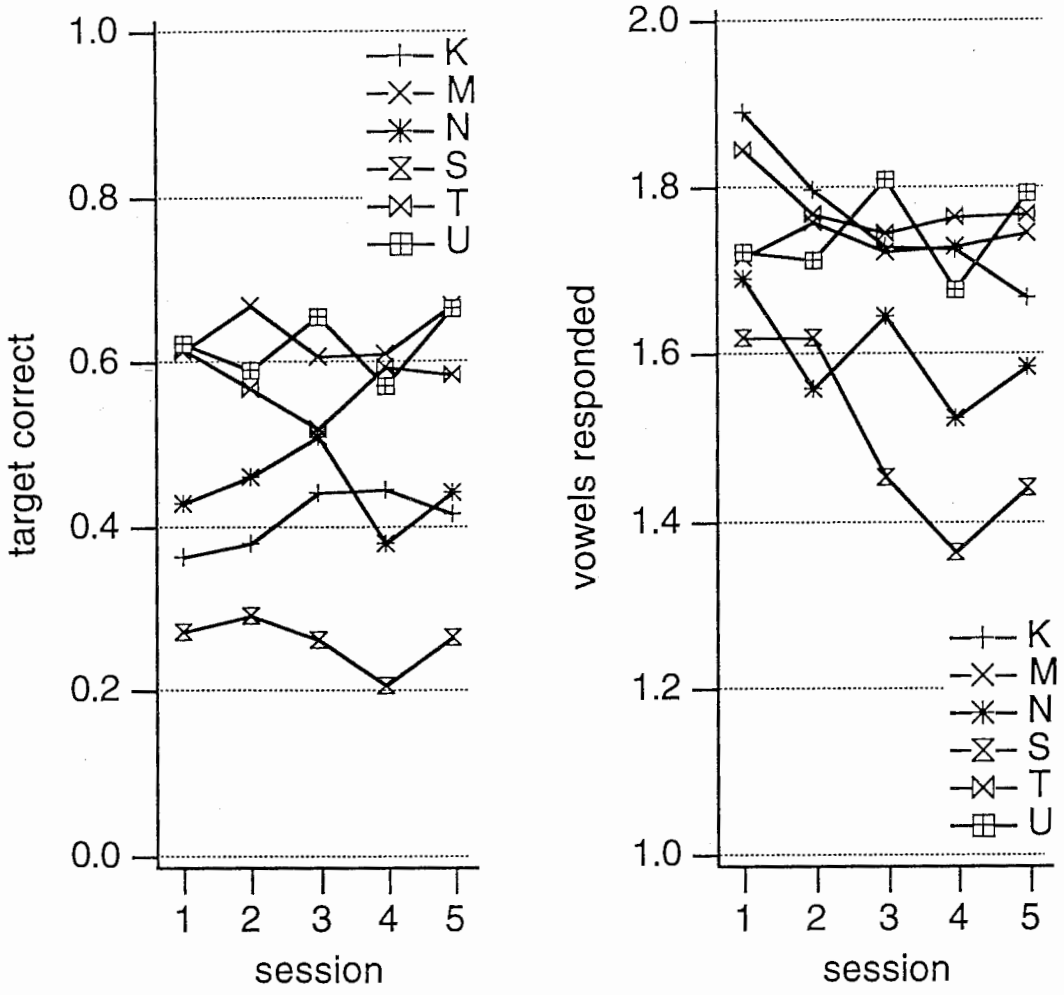


Fig. D-2. Identification rate (left) and number of vowels responded (right) as a function of session number in Experiments 2-4, for all subjects.

Experiments 5-7

Fig D-3. shows the variation in identification rate over sessions. We expected that the new task would induce strong learning effects; this was not the case.

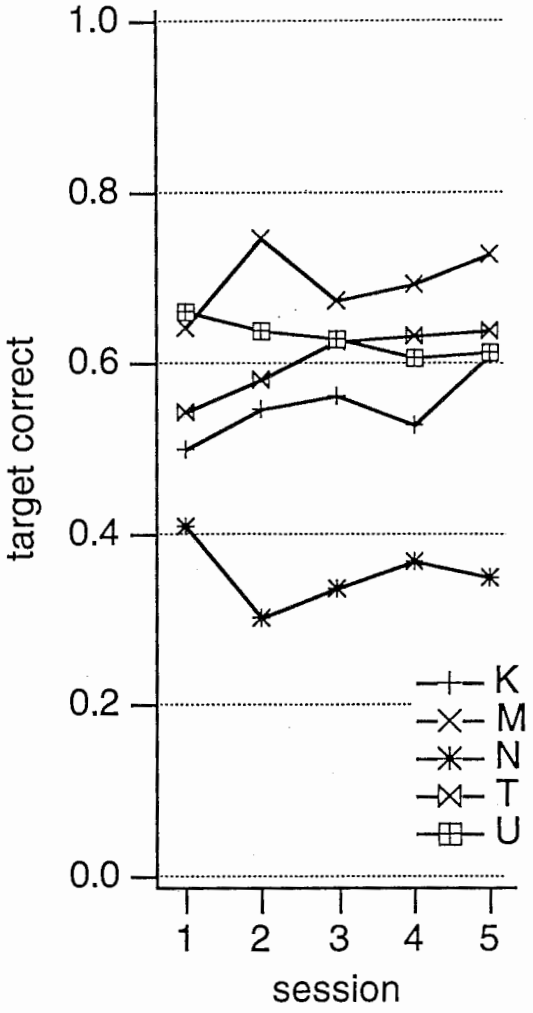


Fig. D-3. Identification rate as a function of session number in Experiments 5-7 for all subjects.