TR－A－0084

Psychoacoustic evidence for the contextual effect model

Masato Akagi

1990. 6.28

ATR視聴覚機構研究所

〒619-02 京都府相楽郡精華町乾谷 ☎07749-5-1411

ATR Auditory and Visual Perception Research Laboratories

Inuidani, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

Telephone: +81-7749-5-1411
Facsimile: +81-7749-5-1408
Telex: 5452-516 ATR J

Psychoacoustic evidence for the contextual effect model

Masato Akagi

ATR Auditory & Visual Perception Research Laboratories
Inuidani, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

Abstract

In previous work towards speech recognition (Akagi, 1989), a model was developed which predicted target formants in reduced vowels based on the interaction between spectral peak pairs. To substantiate this model, two psychoacoustic experiments were carried out which measured the amount of phoneme boundary shift with (1) a single formant stimulus as a preceding anchor and (2) a vowel as a preceding anchor. In the first experiment, a perceptual boundary shift with a single formant anchor was observed. When the results were compared with the spectral peak interaction obtained from real speech data using the model, this comparison showed that the perceptual boundary shift with a single formant anchor is similar to the spectral peak interaction analyzed by the model. Thus, the contextual effect between single formant stimuli should play an important role in phoneme neutralization recovery, and the neutralization recovery model is formulated as the sum of the contextual effects resulting from interaction between spectral peaks. Additionally, a comparison of these results with those of the second experiment showed that the phoneme boundary shift with a vowel anchor can be postulated as the sum of the shift with the single formant anchor and a factor from the preceding anchor. The factor can be estimated by subtracting the sum of the phoneme boundary shifts with the single formant anchors estimated by the model from the boundary shift with a vowel anchor. The difference was represented as a function of the distance between the preceding vowel anchor and the perceived vowel in a phoneme space.

Note : This work has been presented in the 119th meeting of the Acoustical Society of America on May 21-25, 1990 and this paper has been handled attendance of the meeting through the meeting paper-copying service.

1. INTRODUCTION

Analysis of continuous speech reveals that incomplete articulation neutralizes phonemes. This often causes incorrect automatic speech recognition results and is one of the most serious problems in continuous speech recognition. However, upon hearing continuous speech, each phoneme neutralized by co-articulation is perceived as if it were uttered clearly without neutralization. The phenomenon can be explained by a compensation mechanism which presumably exists in the speech perception mechanism (Lindblom and Studdert-Kennedy, 1967). If this compensation mechanism can be modeled, it should be applicable to co-articulation recovery in speech signal processing, particularly in continuous speech recognition.

A lower level contextual effect model as a compensation mechanism has been proposed (Akagi, 1989). A model was developed which predicted target formants in reduced vowels based on the interaction between spectral peak pairs, assuming that the lower level contextual effect is represented as the sum of the interaction function between each spectral peak pair.

To substantiate this model, in this paper, two psychoacoustic experiments were carried out which measured the extent of the phoneme boundary shift as a function of the anchor frequency and an inter stimulus interval (ISI) with (1) a single formant stimulus as a preceding anchor and (2) a vowel as a preceding anchor.

The results of the first experiment showed that a perceptual boundary shift with a single formant anchor was observed. When they were compared with the spectral peak interaction analyzed by the model, this comparison showed that the perceptual boundary shift with a single formant anchor is similar to the spectral peak interaction analyzed by the model. Thus, the contextual effect between single formant stimuli should play an important role in phoneme neutralization recovery, and the model can be formulated as the sum of the contextual effects resulting from interaction between spectral peaks.

Additionally, the comparison of these results with those of the second experiment showed that the phoneme boundary shift with a vowel anchor can be postulated as the

sum of the shift with the single formant anchor and a factor from the preceding anchor. This factor can be estimated by subtracting the sum of the phoneme boundary shifts with the single formant anchors estimated by the model from the boundary shift with a vowel anchor. The difference was represented as a function of the distance between the preceding vowel anchor and the perceived vowel in a phoneme space. It is possible to formulate an additional new model which shifts reference spectral patterns as a function of the distance between preceding spectral patterns and reference spectral patterns.

2. OUTLINE OF THE MODEL

The model is developed based on a relationship between two spectral peaks in the time-frequency domain and is formulated as follows:

$$d(t,f) = \frac{1}{2N+1}\sum_{n=t-N}^{t+N} \frac{1}{K_n} \sum_{m=1}^{K_n} g(|t-n|, p_{nm}-f) \qquad (1)$$

Figure 1 illustrates the concepts of Eq. (1). The peak at time t and frequency f is influenced by another peak p_{nm}. The total interaction is the sum of the interaction between the peak at (t, f) and the peak at (n, p_{nm}) in the scope, $t-N \leq n \leq t+N$ and $m \leq K_n$. Each parameter in Eq. (1) has the following meaning:

g(t,f) :

The function g(t,f) represents interaction between two spectral peaks. The relationship between two peaks is provided only by the differences, |t-n| and p_{nm}-f. If $g(t,f) > 0$, then one peak influences the other to move to a higher frequency.

d(t,f) :

This is the frequency difference between the real spectral peak and its target at time t and frequency f. The function d(t,f) is represented as the sum of the interaction functions g(t,f). If the peak is moved to overshoot d(t,f), it corresponds to its own target.

t and f :

t is time and f is the peak frequency in the bark scale (Zwicker, 1980). This is because g(t,f) is the function of the differences in the time and frequency axes, and the spectrum shift along the bark scale does not influence the accuracy of vowel perception (Hirahara, 1988).

K_n :

Number of peaks at time n.

p_{nm} :

Frequency of the m-th peak at time n, $n \leq K_n$.

When overshooting real peaks, the interaction function g(t,f) has been provided and the parameters, t, n, K_n, and p_{nm}, are obtained from an input spectrum sequence. d(t,f) was calculated by using Eq. (1) and added to a real spectral peak frequency to move to correspond its own target.

3. PURPOSE OF THE EXPERIMENTS

In the model, the interaction function between a spectral peak pair plays an important role as shown in Eq. (1). This can be formulated as a contextual effect resulting from interaction between two single formants in the auditory mechanism. To observe the contextual effects and to substantiate this model, the AX method was used for the following two experiments.

Experiment 1 measured the extent of the phoneme boundary shift with a single formant stimulus as a preceding anchor, when A is a single formant stimulus and X is a vowel. The results of Ex. 1 were compared with the results of the spectral peak interaction obtained from real speech data by using the model, and the relationship between the two results was discussed.

When applying the model to a speech recognition preprocessor, the phoneme environment, e.g. what the adjacent phonemes are, has to be considered. Experiment 2 measured the extent of the phoneme boundary shift with a vowel as a preceding anchor, when both A and X are vowels. The results of Ex. 2 were compared with the results of the contextual effect with a single formant stimulus, and the relationship between the contextual effect with a single formant stimulus and that with a vowel were also discussed.

4. EXPERIMENT CONDITIONS

Two females (YK and HK) who were employed specifically for these experiments served as subjects. Both subjects were native speakers of Japanese with no history of a hearing or speech disorder.

Two sets of stimuli (single formant stimuli and vowels) were synthesized for the stimuli for the preliminary experiment and Exs. 1 and 2, by using a Klatt formant synthesizer with pitch = 140 Hz, 20 kHz sampling, and duration = 200 ms. A 50 ms stimulus was cut with rise and decay times of 10 ms from a 200 ms synthesized sound, as shown in Fig. 2.

(1) single formant sound

20 single formant stimuli were synthesized. The center frequency rose in 1.0 bark steps from 1.0 to 20.0 bark. The bandwidth of all stimuli is set at 50 Hz. Table 1 shows the center frequency of each single formant stimulus. The equation reported by Zwicker (1980) was used for Hz-bark transformation.

(2) vowel sound

85 vowels which varied from Japanese vowel /u/ through /a/ were synthesized. Table 2 provides the frequencies of the five formants and Figure 3 shows the formant positions on the F1-F2 plane.

The single formant and vowel sounds were concatenated to generate stimuli according to the following preliminary experiment and Exs. 1 and 2. The stimuli were randomized and recorded on a DAT (SONY TCD-D10) at 3 second intervals. There was a 1000 Hz, 25 ms pure tone after every 10 trials and a 8 second pause after every 100 trials. The experimental DAT tapes were reproduced on a DAT (SONY TCD-D10) and presented through STAX SR Apro headphones.

Subjects were required to identify each stimulus 20 times, as either the vowel /u/ or /a/. The results were processed to determine the phoneme boundary between /u/ and /a/ by using the SAS PROBIT procedure. The point at which the /a/ judgment exceed 50 % was regarded as the boundary.

5. PRELIMINARY EXPERIMENT

As a preliminary experiment, in order to determine the original phoneme boundary on Japanese /u/-/a/ continua, a single vowel identification test using all of the synthesized vowels as stimuli was repeatedly carried out during the experiment. There were 1700 stimuli (17 F1 × 5 F2 × 20 times) presented as noted in Section 4. The phoneme boundary of each F2 condition was determined by applying the PROBIT procedure. Figure 4 indicates the phoneme boundary of each subject on the F1-F2 plane and ● and ○ represent the performance of subjects YK and HK, respectively.

The results show that the phoneme boundaries almost parallel the F2 axis. This indicates that the subjects judged these stimuli as either the phoneme /u/ or /a/ by using the F1 frequency and that the F2 variation did not influence phoneme perception in this situation. Additionally, Figure 4 illustrates that the boundary frequency does not deviate in time, and that there is an individual difference on the boundary frequency.

6. EXPERIMENT 1

In order to obtain a value of spectral peak pair interaction, a psychoacoustic experiment was carried out which measured the extent of phoneme boundary shift with a single formant stimulus as a preceding anchor. From the preliminary experiment, it can be seen that when the phoneme boundary shifted, the F1 of X, which was a vowel, was influenced by a single formant stimulus.

The stimuli for Ex. 1 were generated by concatenating single formant stimuli and vowels. There were 20 single formant stimuli and 17 vowels, with F2 = 10 bark. The inter stimulus interval (ISI) was varied from 0 ms to 300 ms in 25 ms steps. Thus, there were 88,400 stimuli (20 single formant stimulus × 17 vowels × 13 ISI × 20 times). Figure 5 shows the Ex.1 paradigms.

Figure 6 illustrates the contour lines of the phoneme boundary shift with a single formant stimulus anchor. The vertical axis indicates the difference when the center frequency of the single formant stimulus is subtracted from the original phoneme boundary of each subject, and the horizontal axis indicates the ISI. Additionally, the hatched area shows when the phoneme boundary was shifted to a higher frequency, that is, the F1 of the vowel was perceived as a lower frequency. Thus, when the single formant center frequency was higher than the original phoneme boundary, the white area in the figure indicates that an assimilation effect between a single formant stimulus and the F1 of a vowel is observed.

The following results are from Fig. 6.

(1) A perceptual boundary shift with a single formant anchor was observed.

(2) There is little individual difference between the two subjects' results. This is shown by the comparison of the positions of local maximum and local minimum and their values.

(3) As a variation in the contextual effect along the ISI, an assimilation effect is evident before 70 ms and a contrast effect is evident after 70 ms. Additionally, there is a repetition in the 140 ~ 150 ms period. This period length should be related to the mean syllable length of Japanese.

(4) The contextual effect between single formant stimuli was almost symmetrical where the vertical axis value is 0 and the local maximum and local minimum were placed at regular intervals on the bark scale.

7. ANALYSIS OF REAL SPEECH DATA

To substantiate the model, in this section, the interaction between single formant stimuli illustrated in Fig. 6 and spectral peak interaction analyzed real speech data by using the model were compared.

The database used in calculation of the interaction between a spectral peak pair was as follows: 226 different Japanese words, including (/a/,/i/,/u/,/e/,/o/ = (189, 156, 144, 115, 124) vowels and an uncontrolled consonant environment, which were uttered by one male speaker. In each word, three spectral peaks were labeled in vowel intervals and all spectral peaks were labeled in consonant intervals. The target peak is adopted for the spectral peak mean computed for vowels uttered in isolation. d(t,f) in Eq. (1) sets the difference between a real spectral peak and its target frequency.

In order to determine the interaction function g(t,f) from the difference d(t,f) between a real spectral peak and its target by using Eq. (1), a pseudo inverse matrix like the design of a 2-dimensional filter was employed because the coefficient matrix of the linear equations is singular.

A scope for calculating the interaction function g(t,f) is -300 ms $\leq$ t $\leq$ 300 ms and -17 bark $\leq$ f $\leq$ 17 bark because the scope for t must cover more than 3 syllables and the scope for f must cover the maximum difference between two spectral peaks. The scope was shifted $\pm$50 ms at the vowel center to enlarge the rank of the coefficient matrix.

Figure 7 shows the analyzed interaction function g(t,f) obtained from d(t,f) of real speech data. Figure 7 was modified to make it easy to compare with Fig. 6 which displayed the frequency distance between two peaks from -5 bark to 15 bark, and hatched in the same way as Fig. 5, and g(t,f) was extrapolated in 0 ~ 10 ms.

The comparison of the positions of local maximum and minimum in Fig. 6 with those in Fig. 7 indicates that the frequency interval is similar and that repeated patterns in time are also similar. The values of local maximum and minimum in Fig. 7 are three times larger than those in Fig. 6. However, the result in Fig. 7 was calculated under the assumption that a real spectral peak must correspond to its target. Thus, the aspects of the two figures are similar.

These results suggest that the contextual effect between single formant stimuli should play an important role in phoneme neutralization recovery and that the model can be formulated as the sum of the contextual effect resulting from interaction between spectral peaks.

8. EXPERIMENT 2

In order to observe compare the contextual effect with a single formant anchor and with a vowel anchor, the extent of the phoneme boundary shift with a vowel as a preceding anchor was measured.

The stimuli for Ex. 2 were generated with concatenating vowels. There were 6 vowels (F1 = 3.0, 4.2, 5.0, 5.6, 6.4, 7.4 bark) for the preceding anchor and 17 vowels for the perceived vowels. The F2 of all stimuli was 10 bark. The inter stimulus interval (ISI) was varied from 25 ms to 300 ms in 25 ms steps. Thus, there were 24480 stimuli (6 vowels × 17 vowels × 12 ISI × 20 times). Figure 8 shows the Ex. 2 paradigms.

Figure 9 (a) shows the results for subject YK with a preceding anchor. The vertical axis indicates an assimilation. If the results below assimilation = 0 are noted, it suggests that a contrast effect is observed. Thus, when a preceding anchor is far from the original phoneme boundary (3.0, 6.4, 7.4 bark) there is a large contrast effect, otherwise there is a small contrast or assimilation effect. Since the results of Ex. 2 which include interactions between single formant pairs will be considered, let us subtract the sum of the phoneme boundary shifts with the single formant anchors estimated by the model using Ex. 1 results from the boundary shift with a vowel anchor.

Figure 10 shows the differences and their fitted linear lines. These could represent a new factor for phoneme boundary shift when the preceding anchor is a vowel. Additionally, Figure 11 shows values of fitted linear functions at ISI = 200 ms as a function of the difference between the original phoneme boundary and the F1 of a preceding anchor.

In Figs. 10 and 11, the results when a preceding anchor is far from the phoneme boundary show a contrast effect, and the results when an anchor is close to the phoneme boundary show an assimilation effect. The graphs in Fig. 11 were almost symmetrical where the difference was zero. Thus, this factor is represented as a function of the distance between a preceding anchor and a perceived vowel in a phoneme space. The function should be formulated like a function representing a lateral inhibition.

A contextual effect model reconsidered according to two psychoacoustic experiments in order to apply automatic speech recognition, is represented as shown in Fig. 12. First, an input spectrum sequence goes through a model that predicts target spectral peaks in reduced vowels based on the interaction between spectral peak pairs. Next, the reference spectral patterns shift to pull back reduced spectral patterns into each of their correct categories based on the interaction between a preceding spectrum sequence and reference spectral patterns in a spectral pattern space.

9. CONCLUSION

In order to substantiate the model, two psychoacoustic experiments were carried out which measured the extent of phoneme boundary shift with (1) a single formant stimulus as a preceding anchor and (2) a vowel as a preceding anchor. The results of Ex. 1 showed that:

(1) the perceptual boundary shift with a single formant anchor was observed,

(2) the perceptual boundary shift with a single formant anchor is similar to the spectral peak interaction analyzed by the model,

(3) the contextual effect between single formant stimuli should play an important role in phoneme neutralization recovery, and

(4) the neutralization recovery model is formulated as the sum of the contextual effects resulting from interaction between spectral peaks.

Additionally, a comparison of the results of Ex. 1 and Ex. 2 showed that:

(5) the phoneme boundary shift with a vowel anchor can be postulated as the sum of the shift with the single formant anchor and a factor from the preceding anchor,

(6) the factor can be estimated by subtracting the sum of the phoneme boundary shifts with the single formant anchors estimated by the model from the boundary shift with a vowel anchor, and

(7) the difference was represented as a function of the distance between the preceding vowel anchor and the perceived vowel in a phoneme space.

These results show it is possible to formulate an additional new model which shifts reference spectral patterns as a function of the distance between preceding spectral patterns and reference spectral patterns.

Acknowledgement

A portion of this study was carried out during the author's stay at MIT. He wishes to thank Dr. Victor Zue at LCS, MIT for his kind help.

References

Akagi, M. (1989). "Evaluation of spectrum target prediction model in speech perception", J. Acoust. Soc. Am. Suppl. 1, 85, II8.

Hirahara, T. (1988). "On the role of the fundamental frequency in vowel perception", J. Acoust. Soc. Am., Suppl. 1, 84, WW11.

Lindblom, B. E. F. and Studdert-Kennedy, M. (1967). "On the role of formant transition in vowel recognition", J. Acoust. Soc. Am., 42, 4, 686-694.

Zwicker, E. and Terhardt, E. (1980). "Analytic expressions for critical-band rate and critical bandwidth as a function of frequency", J. Acoust. Soc. Am., 68, 5, 1523-1525.

TABLES

Table 1. Single formant stimuli (bandwidth = 50 Hz, F_0 = 140 Hz fixed)

Stimulus number	Bark	Hz
0	1.0	102
1	2.0	204
2	3.0	309
3	4.0	417
4	5.0	531
5	6.0	651
6	7.0	781
7	8.0	922
8	9.0	1079
9	10.0	1255
10	11.0	1456
11	12.0	1691
12	13.0	1968
13	14.0	2303
14	15.0	2711
15	16.0	3212
16	17.0	3823
17	18.0	4554
18	19.0	5413
19	20.0	6414

Table 2. Synthesized vowel stimuli for /u/-/a/ continua (F_0 = 140 Hz)

Formant	Bandwidth	Center frequency	
	Hz	Bark	Hz
F1	50	3.0	309
		3.6	373
		4.0	417
		4.2	439
		4.4	462
		4.6	484
		4.8	507
		5.0	531
		5.2	554
		5.4	578
		5.6	602
		5.8	626
		6.0	651
		6.2	676
		6.4	715
		6.8	754
		7.4	836
F2	70	9.0	1079
		9.5	1164
		10.0	1255
		10.5	1352
		11.0	1456
F3	110	14.4	2450
F4	250	16.2	3300
F5	200	16.9	3750

FIGURE CAPTIONS

Fig. 1. Concepts of the model.

Fig. 2. Truncation window for a single formant and vowel sounds.

Fig. 3. F1-F2 formant positions of synthesized vowels.

Fig. 4. Phoneme boundary on /u/-/a/ continua.

Fig. 5. Experimental paradigm for Experiment 1.

Fig. 6. Phoneme boundary shift with a single formant anchor, (a) subject YK and (b) subject HK.

Fig. 7. Analyzed result of peak interaction function g(t,f).

Fig. 8. Experimental paradigm for Experiment 2.

Fig. 9. Phoneme boundary shift with a vowel anchor, (a) subject YK and (b) subject HK.

Fig. 10. Differences between the sum of the phoneme boundary shifts with the single formant anchors estimated by the model using Ex. 1 results and the boundary shifts with a vowel anchor, (a) subject YK and (b) subject HK.

Fig. 11. Values of fitted linear functions at ISI = 200 ms as a function of the difference between the original phoneme boundary and F1 of a preceding anchor.

Fig. 12. Revised contextual effect model.

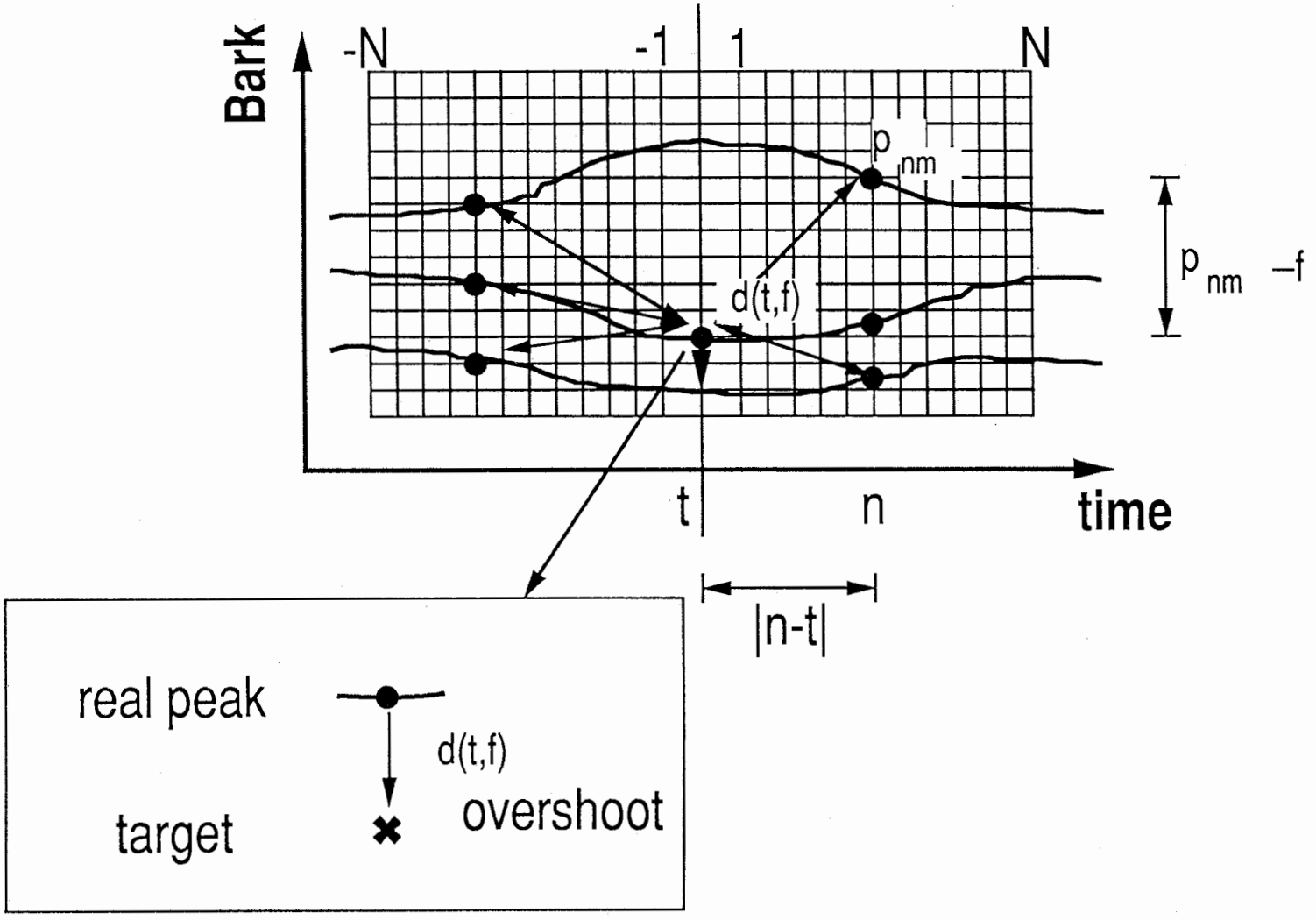

Bark
-N
-1
1
N
p nm
d(t,f)
p nm –f
t
n
time
|n-t|
real peak
d(t,f)
target
overshoot

Fig. 1

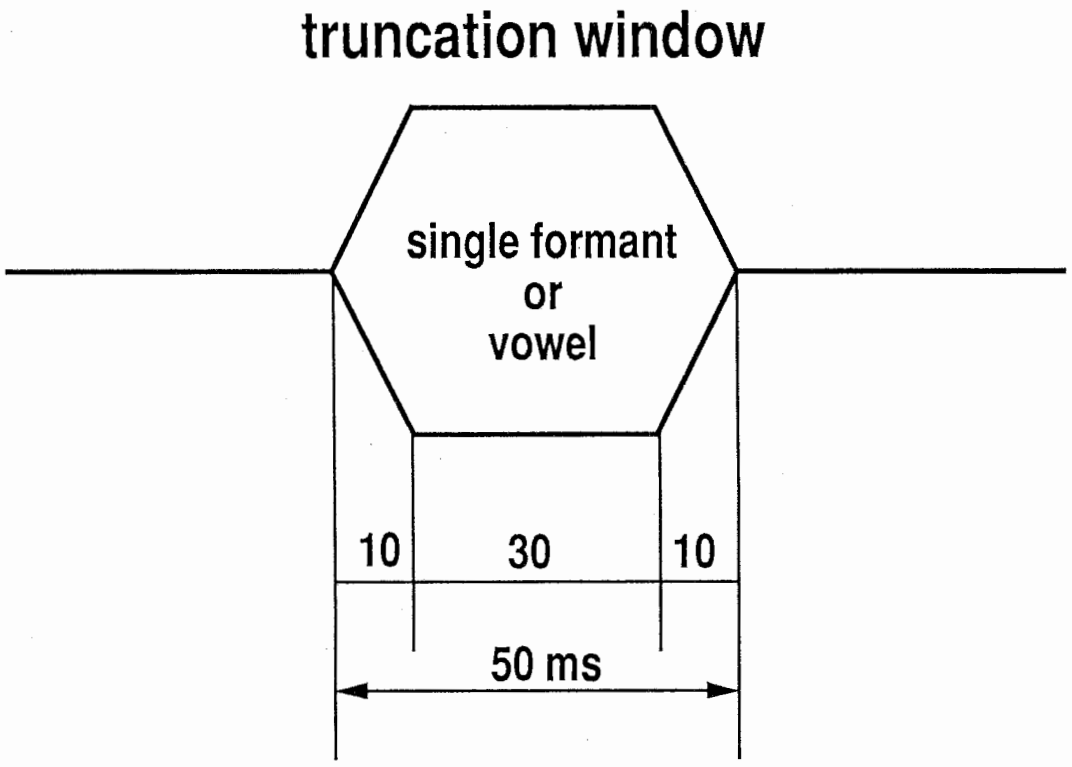


Fig. 2

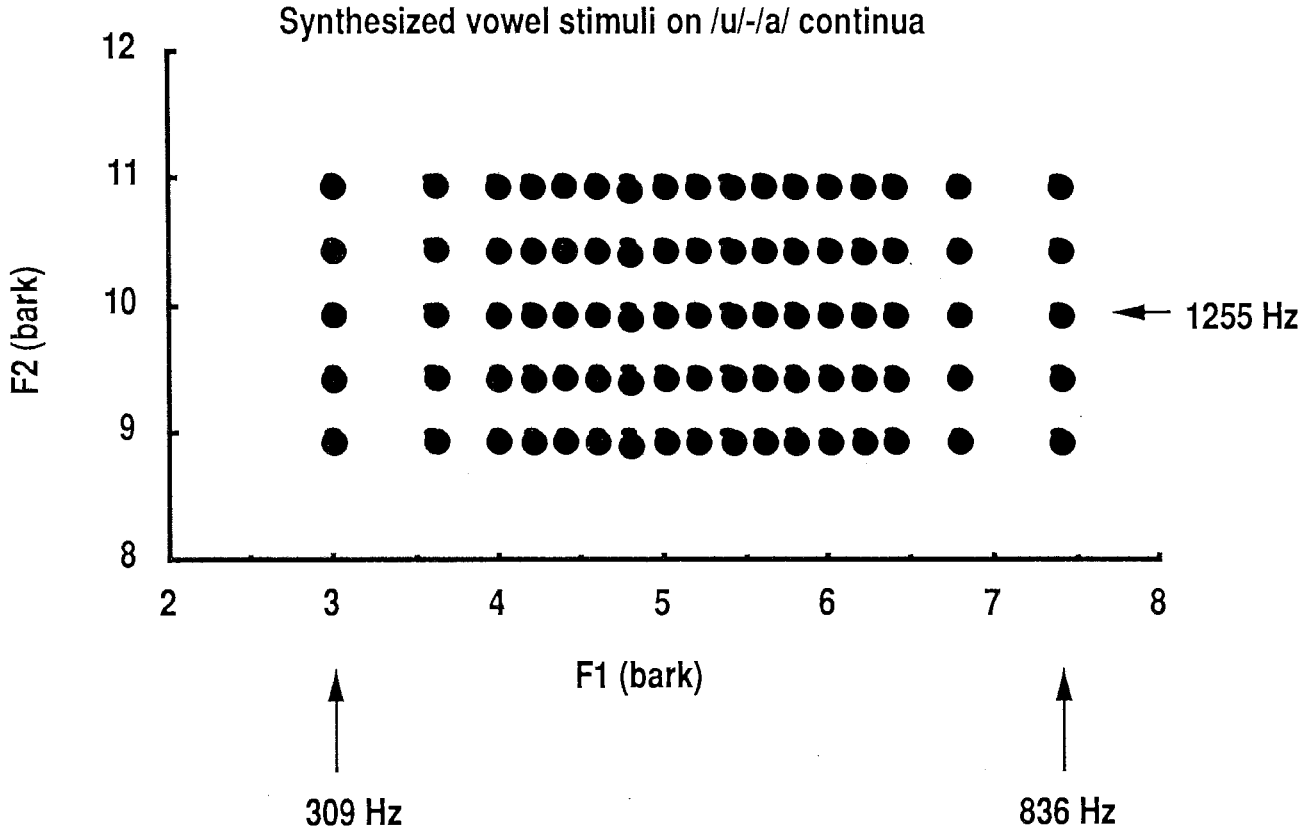

Synthesized vowel stimuli on /u/-/a/ continua
12
11
10
9
8
F2 (bark)
2
3
4
5
6
7
8
F1 (bark)
1255 Hz
309 Hz
836 Hz

Fig. 3

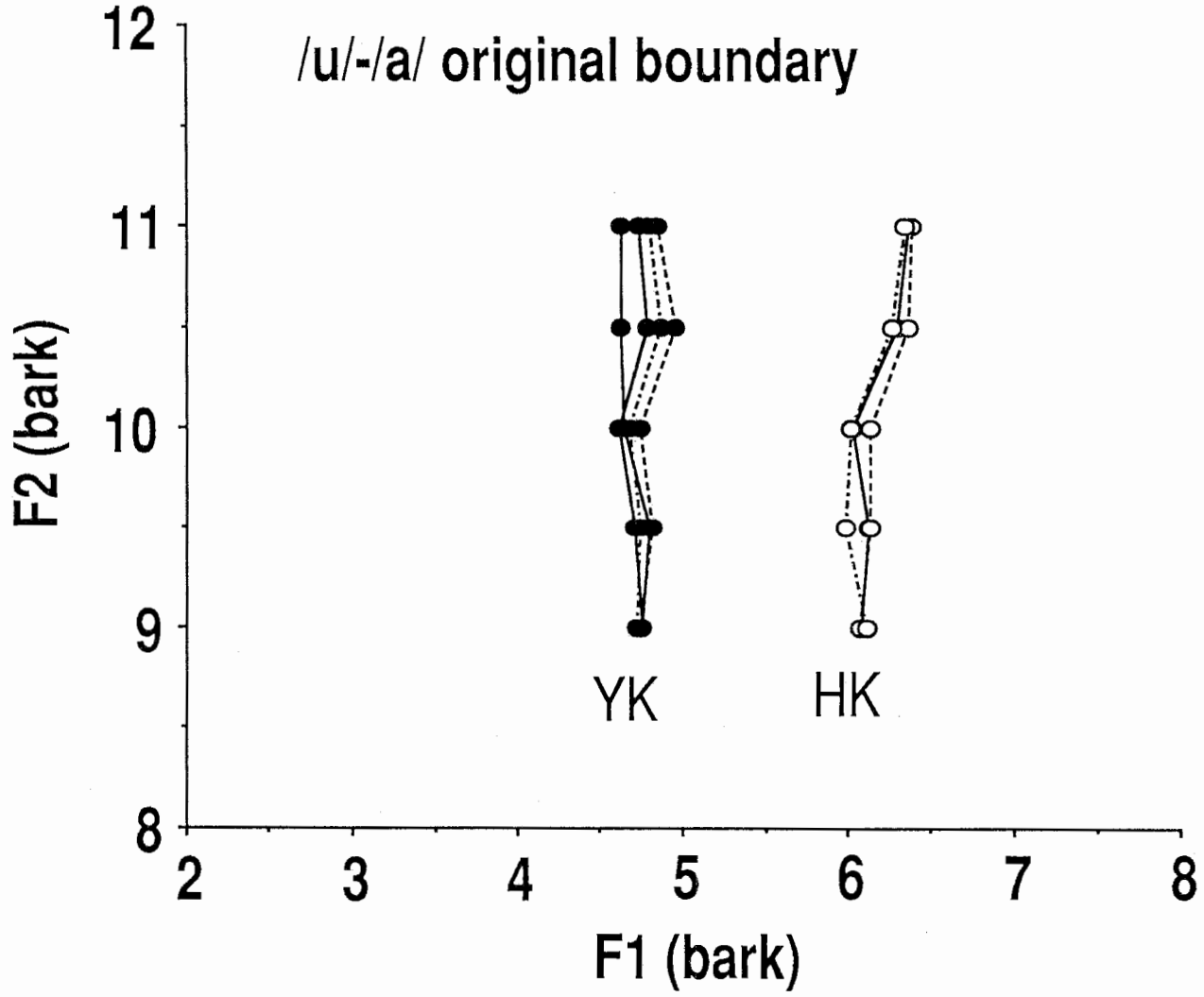


Fig. 4

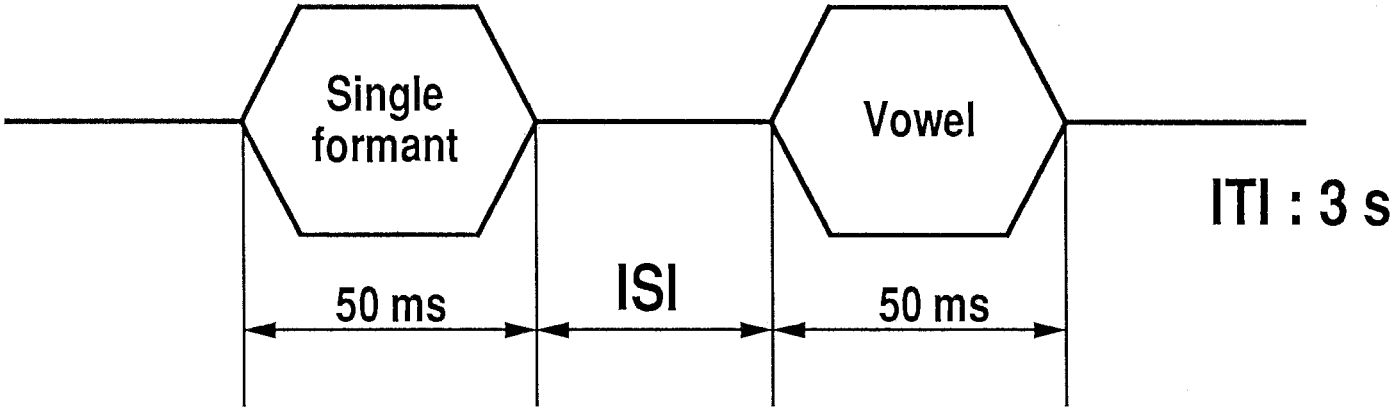

Single formant
Vowel
ITI : 3 s
50 ms
ISI
50 ms
ISI : 0 ~ 300 ms in 25 ms steps

Fig. 5

(a)

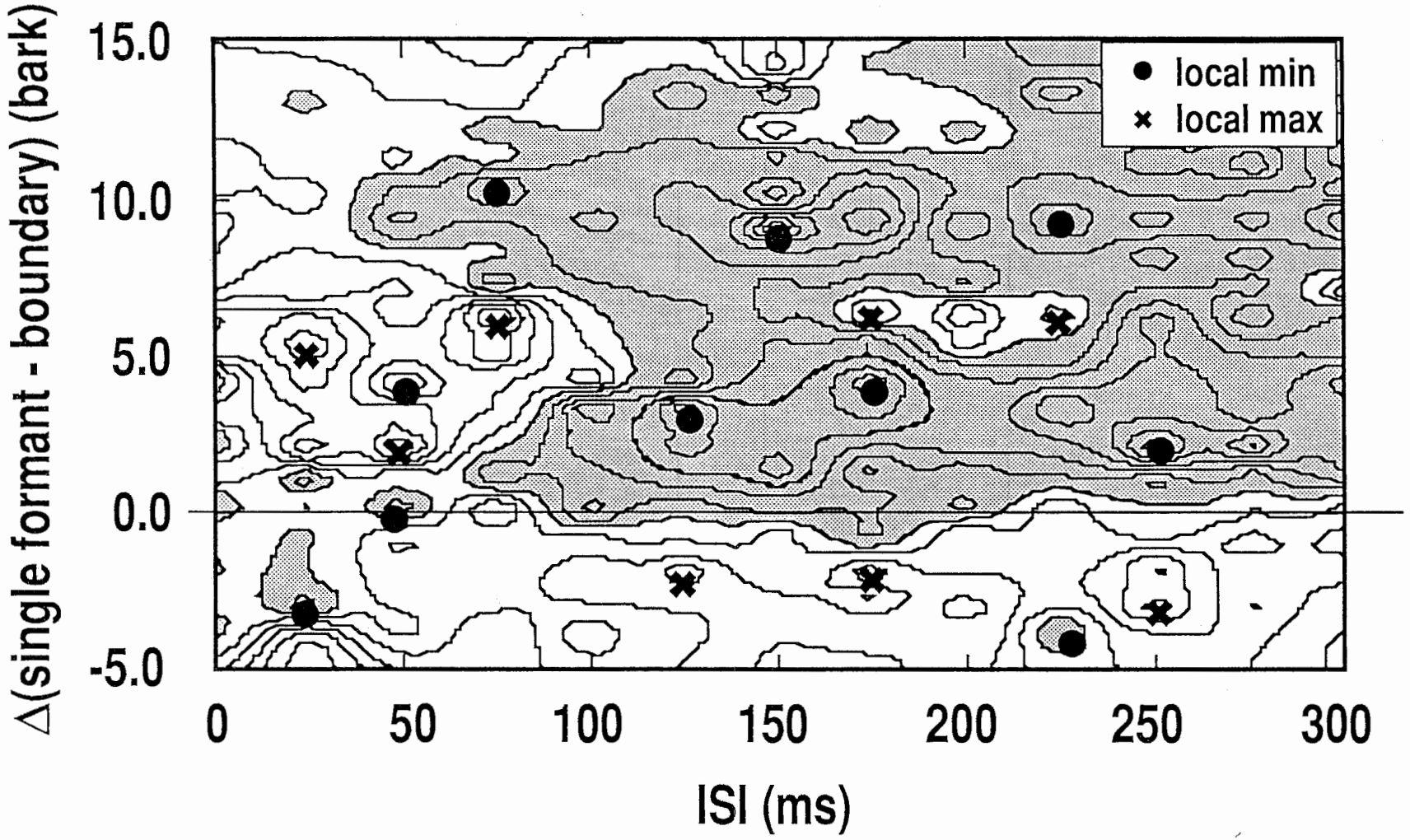

15.0
10.0
5.0
0.0
-5.0
Δ(single formant - boundary) (bark)
0
50
100
150
200
250
300
ISI (ms)
local min
local max

Fig. 6 (a)

(b)

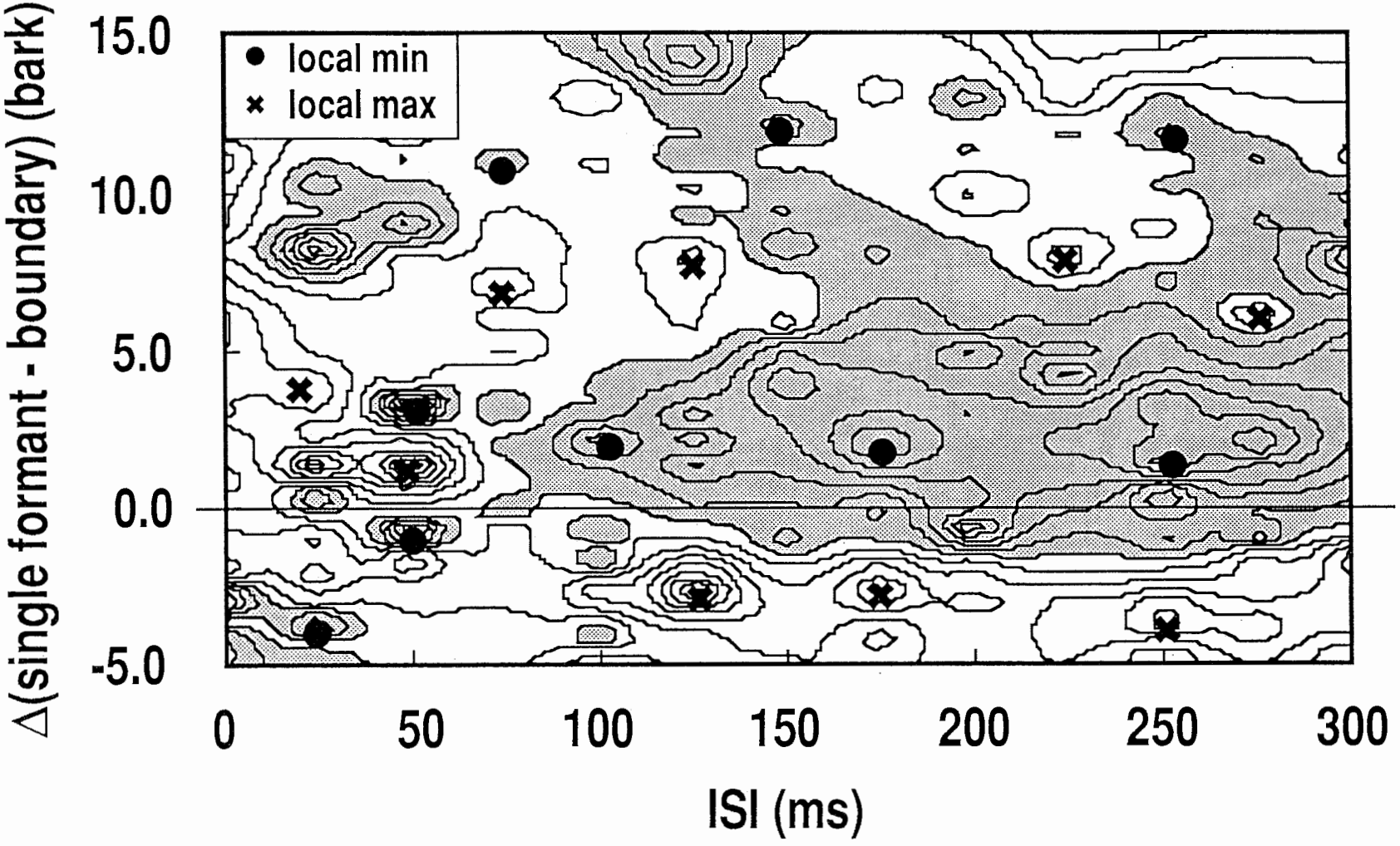

local min
local max
15.0
10.0
5.0
0.0
-5.0
Δ(single formant - boundary) (bark)
0
50
100
150
200
250
300
ISI (ms)

Fig. 6 (b)

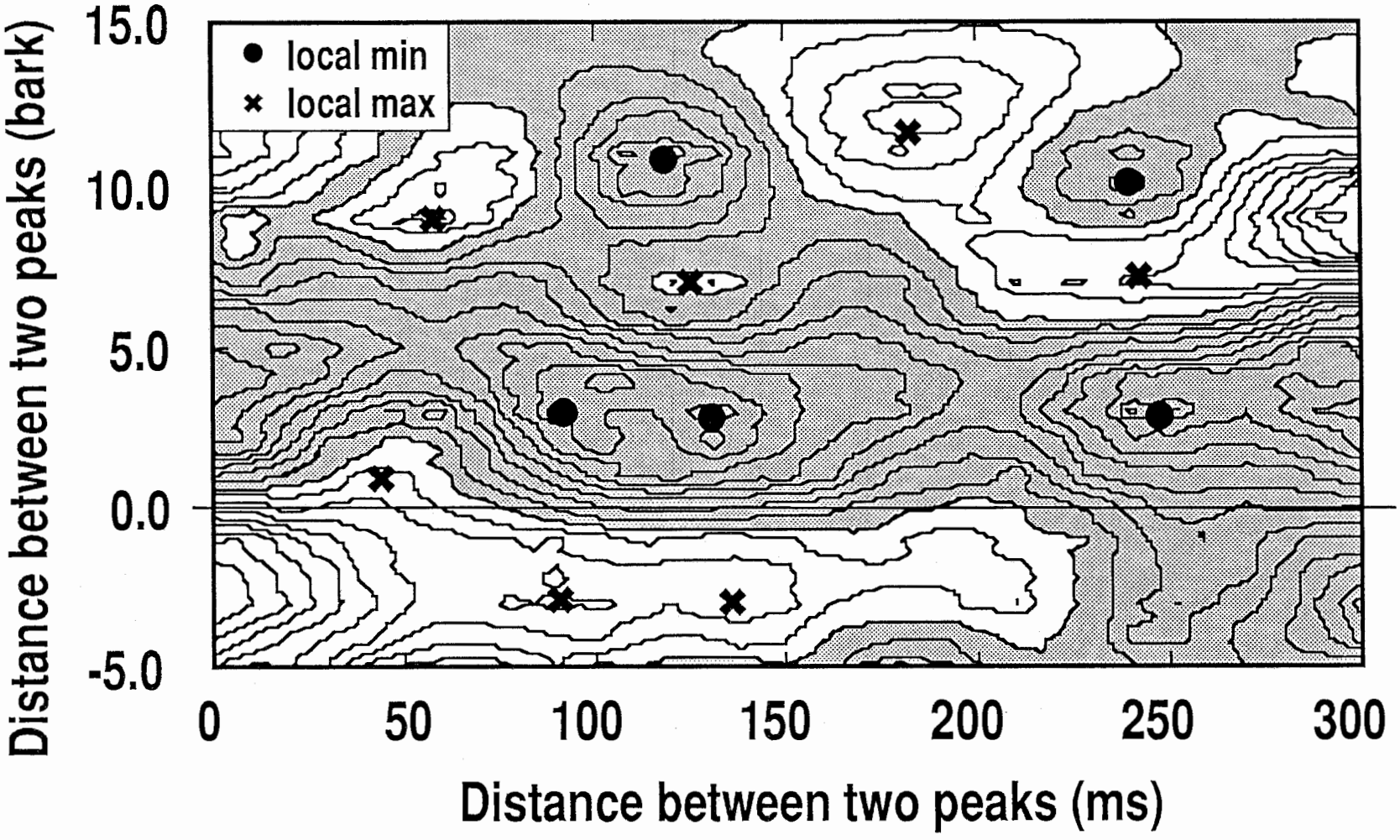
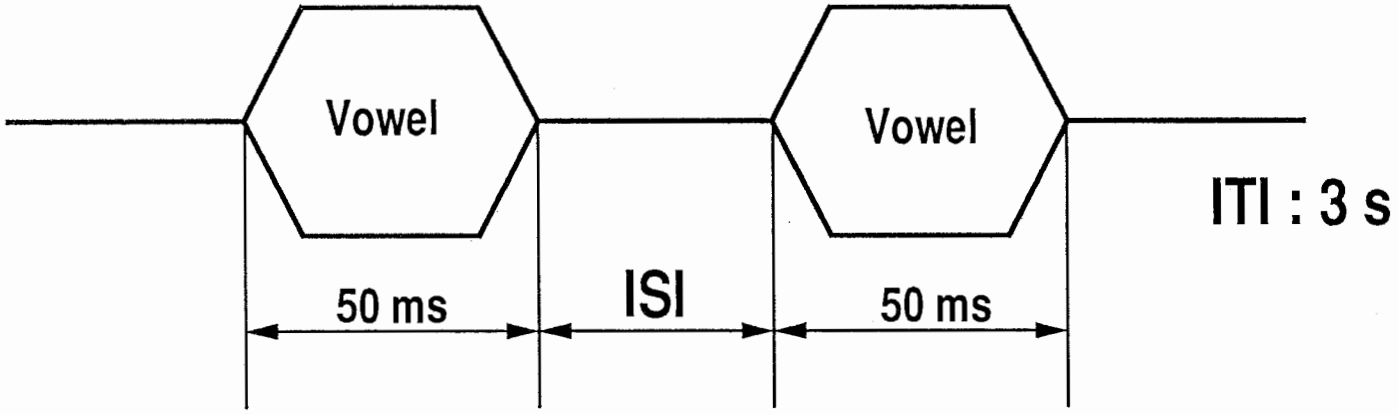

local min
local max
15.0
10.0
5.0
0.0
-5.0
Distance between two peaks (bark)
0
50
100
150
200
250
300
Distance between two peaks (ms)

Fig. 7



ISI : 25 ~ 300 ms in 25 ms steps

Fig. 8

(a)

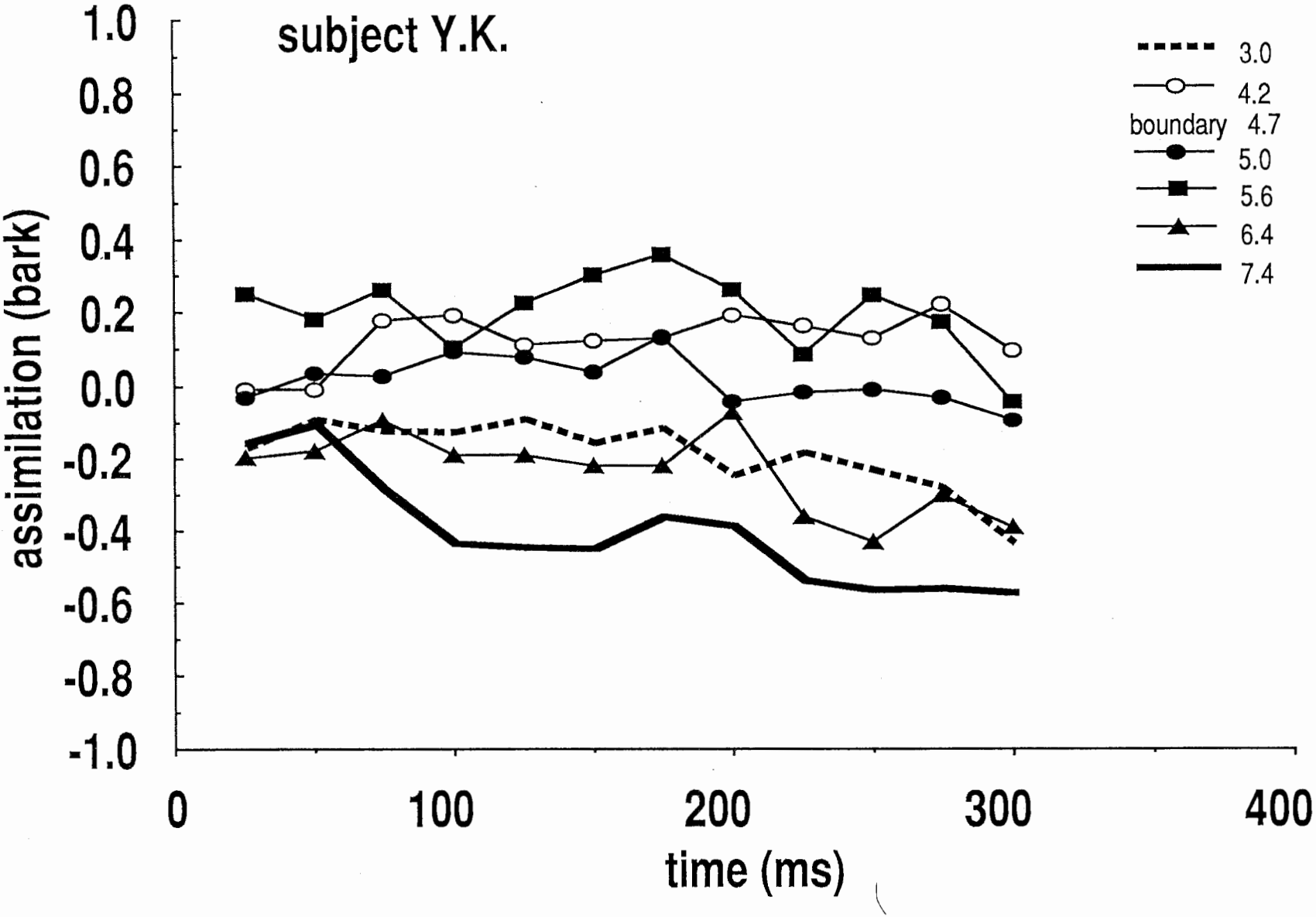

subject Y.K.
assimilation (bark)
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
0
100
200
300
400
time (ms)
3.0
4.2
boundary 4.7
5.0
5.6
6.4
7.4

Fig. 9 (a)

(b)

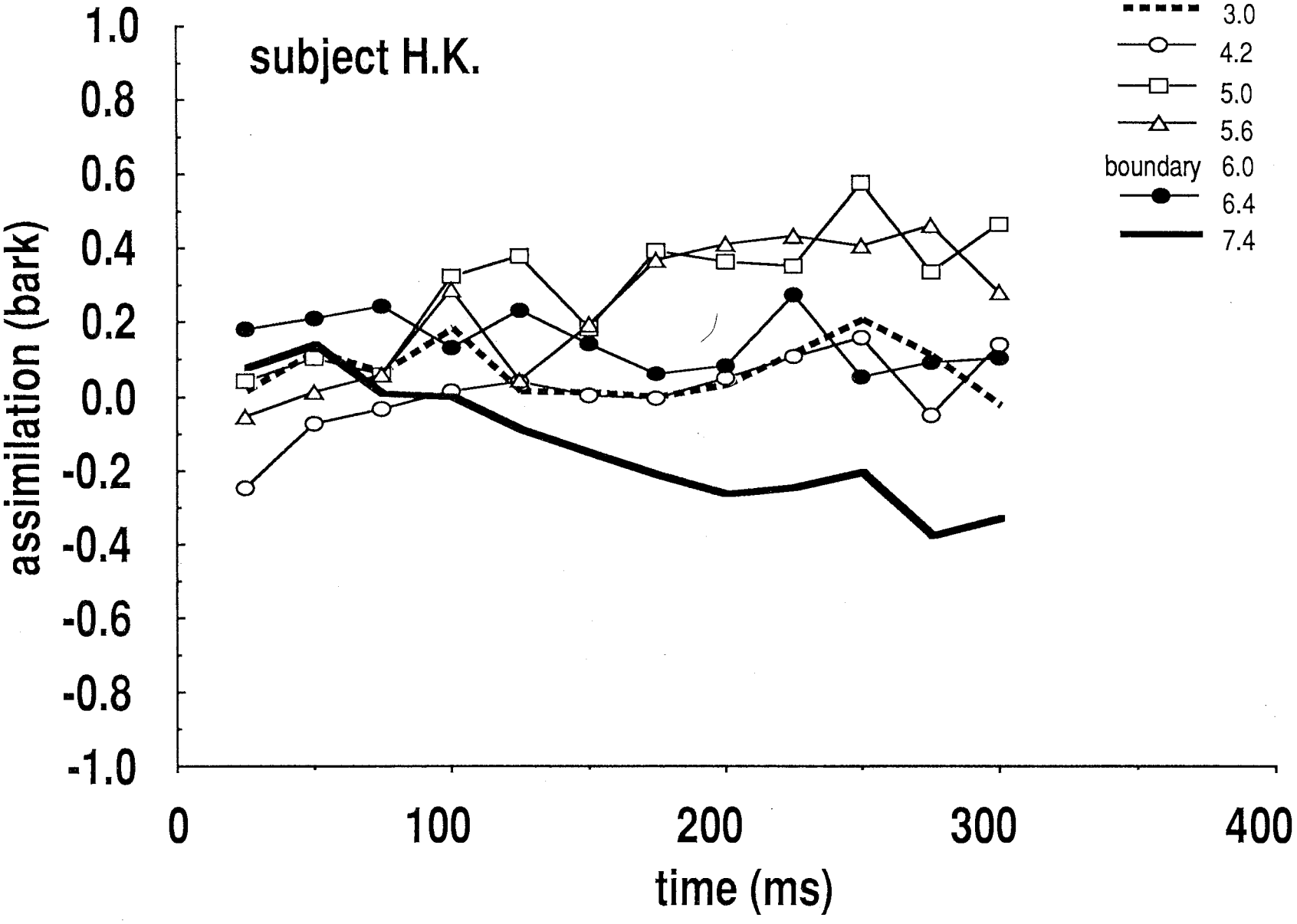

subject H.K.
assimilation (bark)
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
0
100
200
300
400
time (ms)
3.0
4.2
5.0
5.6
boundary 6.0
6.4
7.4

Fig. 9 (b)

(a)

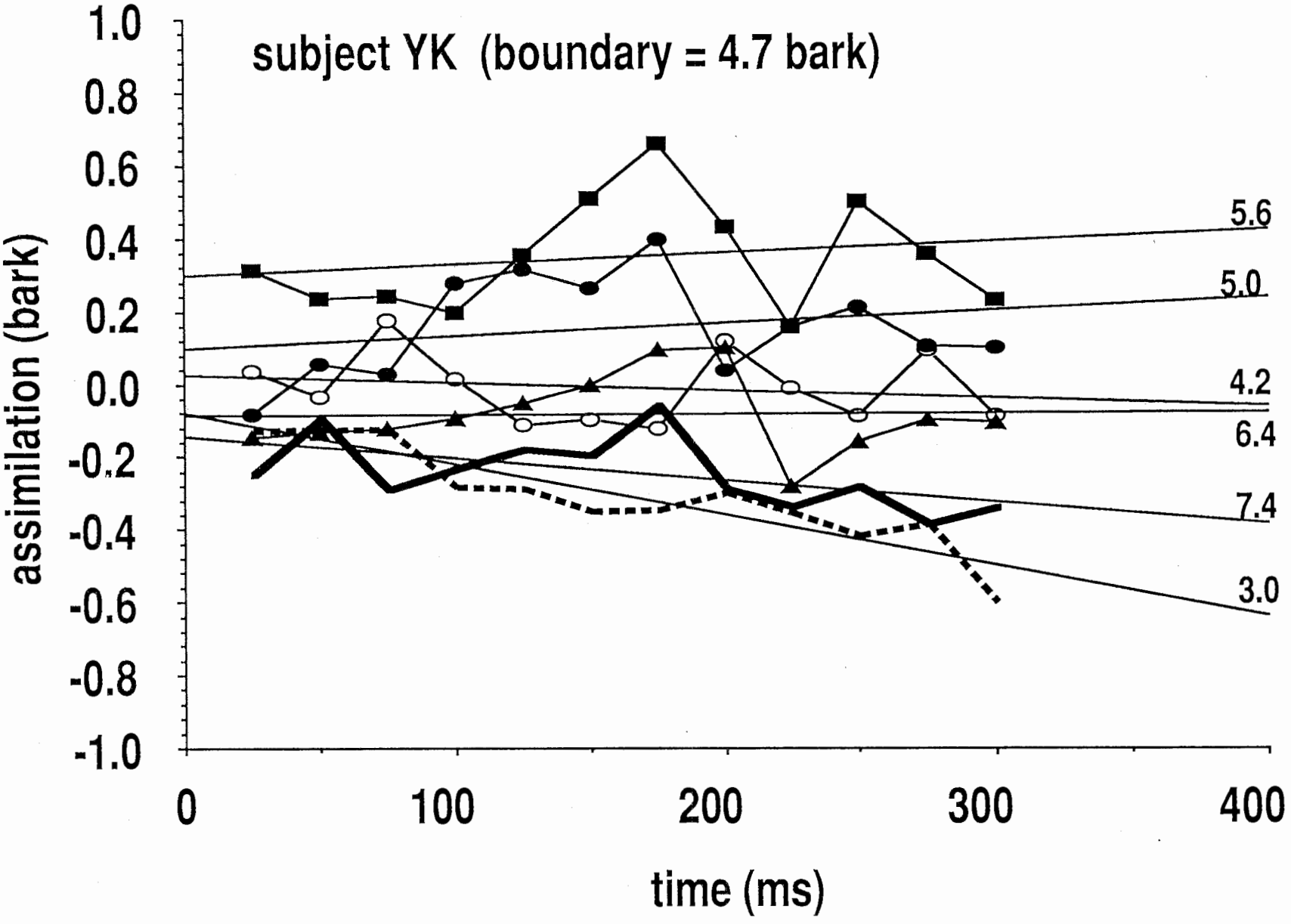

subject YK (boundary = 4.7 bark)
assimilation (bark)
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
0
100
200
300
400
time (ms)
5.6
5.0
4.2
6.4
7.4
3.0

Fig. 10 (a)

(b)

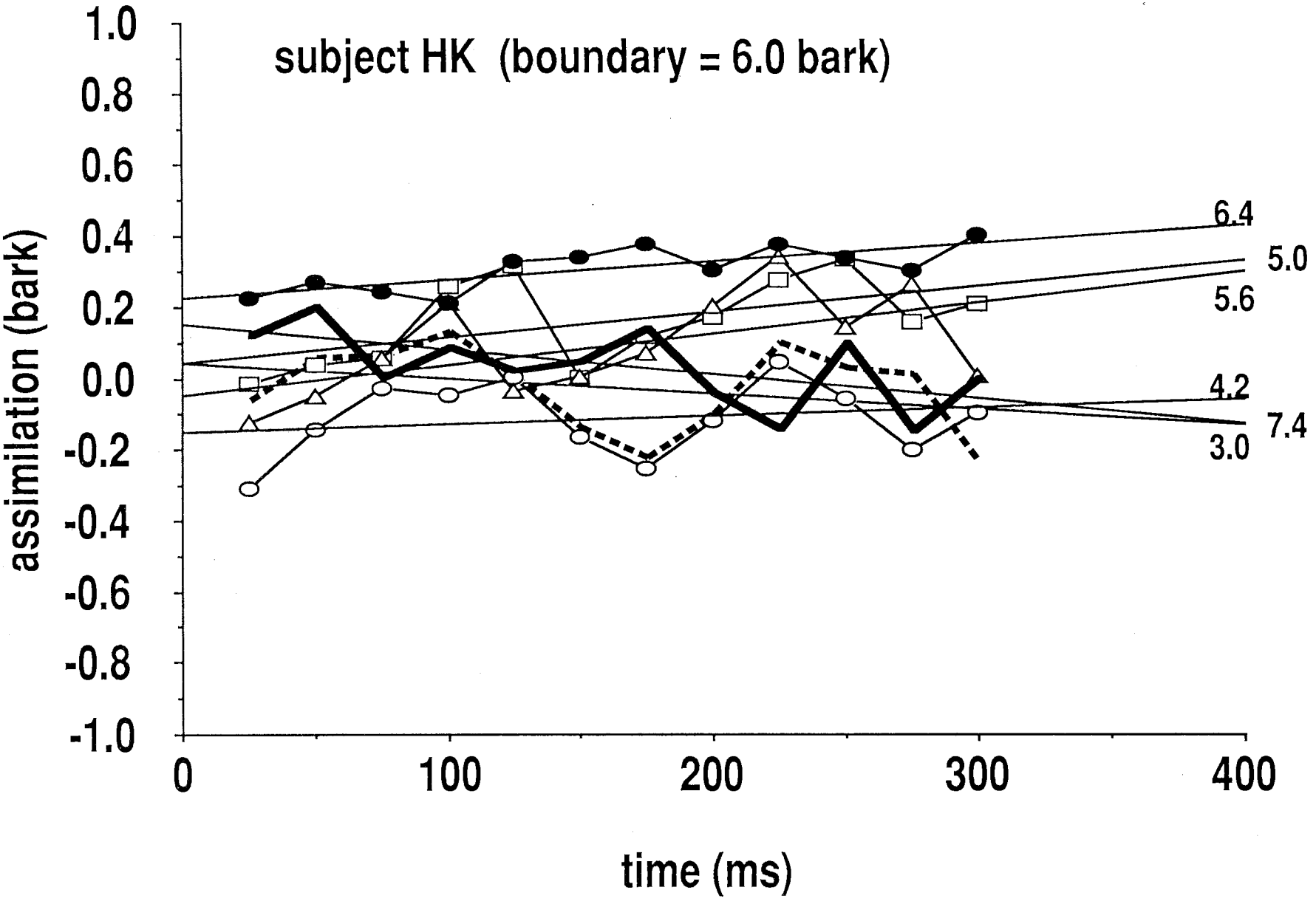

subject HK (boundary = 6.0 bark)
assimilation (bark)
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
0
100
200
300
400
time (ms)
6.4
5.0
5.6
4.2
7.4
3.0

Fig. 10 (b)

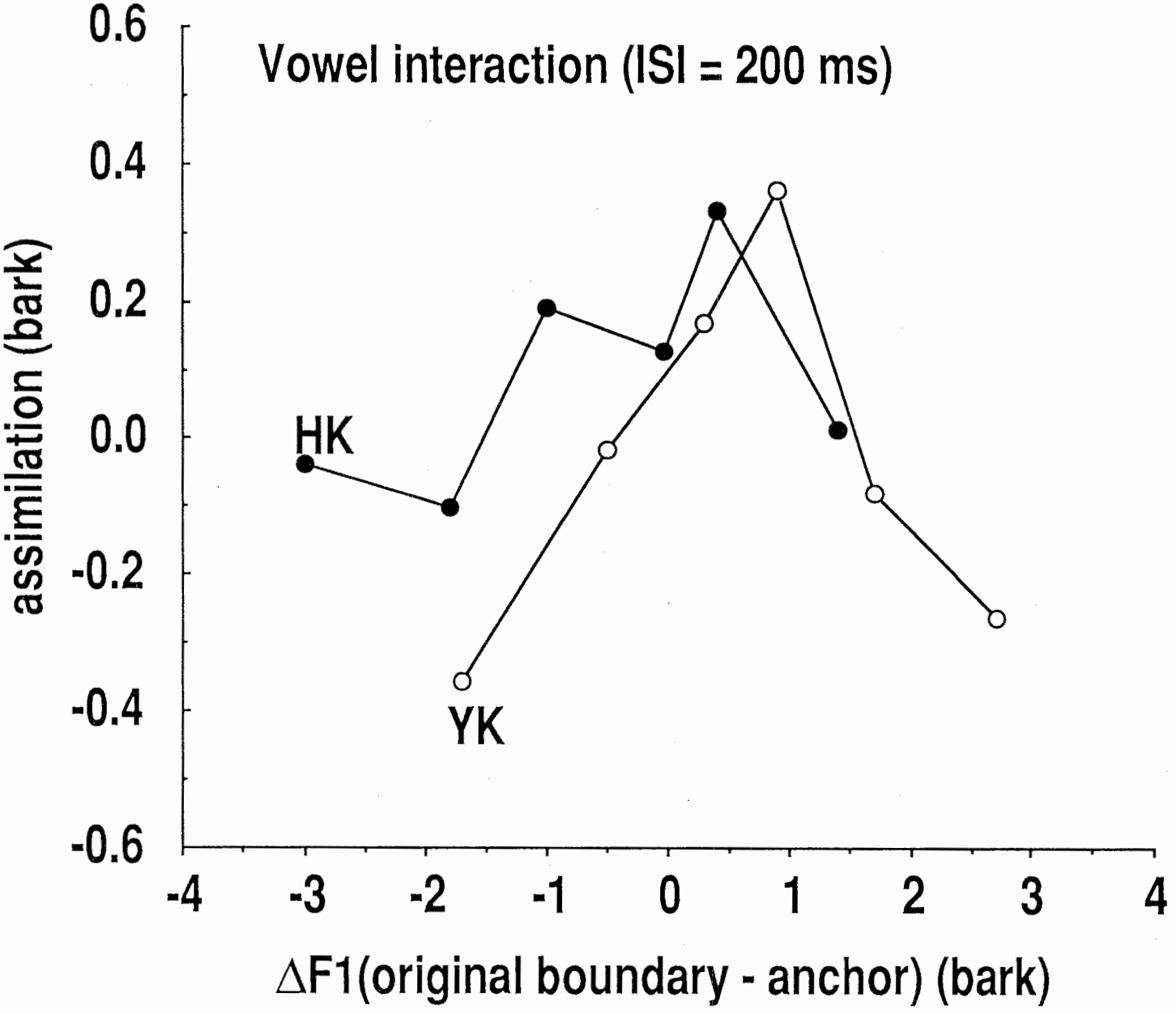

Vowel interaction (ISI = 200 ms)
assimilation (bark)
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
HK
YK
-4
-3
-2
-1
0
1
2
3
4
ΔF1(original boundary - anchor) (bark)

Fig. 11

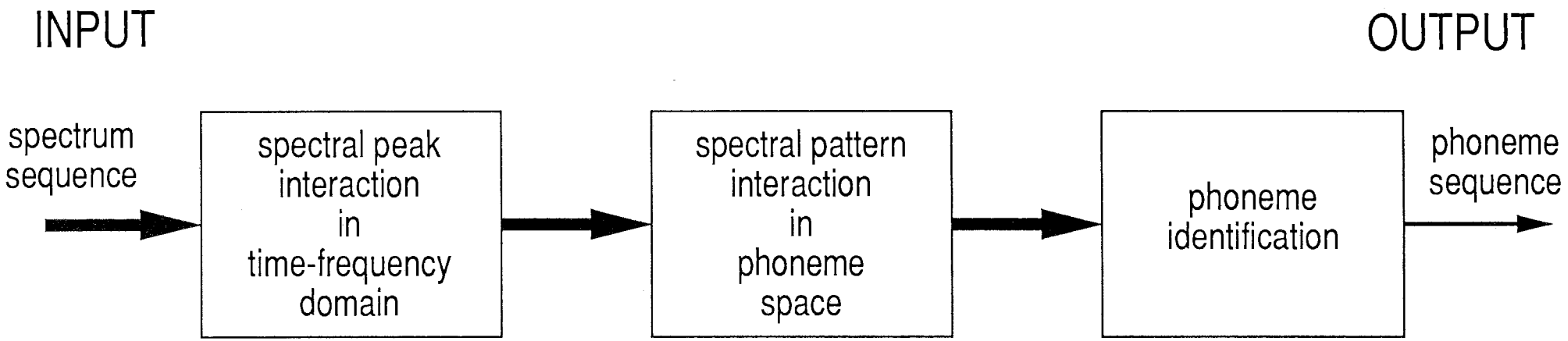

INPUT
OUTPUT
spectrum sequence
spectral peak interaction in time-frequency domain
spectral pattern interaction in phoneme space
phoneme identification
phoneme sequence

Fig. 12