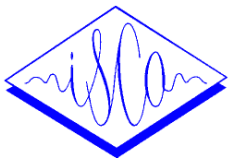

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

PERCEPTION OF CENTRAL VOWEL WITH PRE- AND POST-ANCHORS

Masato Akagi[1], Astrid van Wieringen[2], and Louis C. W. Pols[2]

1) Japan Advanced Institute of Science and Technology, Hokuriku
15 Asahidai, Tatsunokuchi, Ishikawa 923-12, Japan
2) Institute of Phonetic Sciences/IFOTT, Univ. of Amsterdam
Herengracht 338, 1016 CG, Amsterdam, The Netherlands

Abstract

A vowel identification and a vowel matching experiment were performed to examine how preceding and following anchor signals affect central vowel perception. Previous experiments had shown that dynamical aspects of stimuli and the relation between the central vowel and the adjacent anchors may induce overshoot or extrapolation during stimulus processing. The present experiments examine the relative importance of the surrounding steady-state and the formant transitions with regard to center vowel extrapolation. Assuming that continuous phoneme sequences such as consonant-vowel-consonant (CVC) or VVV can be constructed with steady-states, transitions and vowels, four stimulus conditions are used: vowel presented in isolation (called Ref.), vowel with transitions (called Λ), isolated vowel surrounded by steady-states (called Π), and vowel with transitions and steady-states (called Ω). The central vowels were always 5-formant synthesized vowels, whereas the surrounding steady state and the transitions were either single-formant type sounds or 5-formant type sounds. The experimental results suggest that: (1) central vowel extrapolation occurs with Ω-type stimuli, in both single- and 5-formant conditions, whereas averaging effects are observed with Λ- and Π-type stimuli for some of the subjects. The overall order of the amount of overshoot is Ω > Π > Λ > Ref. in the 5-formant condition. The most natural-sounding Ω-type stimuli showed the largest amount of overshoot, and (2) the amount of overshoot with a 5-formant steady-state is larger than with a single-formant steady-state especially for the Π-type stimuli. This might be an indication that the 'vowelness' of the pre- and post-anchors also contributes to the amount of overshoot. The matching results were less consistent.

I. Introduction

Identification of a central vowel surrounded by phonetic context can be affected by pre- and post-phonemes, especially when the vowel is ambiguous. Even though the vowel presented in isolation is perceived correctly, it can be perceived as an another phoneme in contextual situations.

Previous research has shown that the dynamics of the spectrum transition and the relations between the central vowel and adjacent phonemes may induce overshoot or extrapolation in the processing of stimuli.

Lindblom and Studdert-Kennedy[1] and Kuwabara and Sakai[2] indicated that rapidly changing spectral sequences affect the perception of vowels and that the vowels are perceived with some overshoot or extrapolation. Recently, not only overshoot but also averaging is observed in the perception of F1 time-varying trajectories[3][4]. Additionally, previous other studies have also suggested that dynamic spectral features are important for phoneme perception[5]-[7]. These studies investigated the relations between the central vowel and pre- and/or post-transitions.

On the other hand, relations between the central vowel and the adjacent steady-state speech-like/non-speech anchors separated by silent gaps have been also investigated[8]-[11]. These studies have shown that the amount of overshoot and undershoot is related to the time and frequency differences between the central vowel and the anchors and that the amount of overshoot with speech-like anchors is larger than that with non-speech anchors. Although these studies suggested useful information about the perception of vowels surrounded by phonetic context, the situation is still unnatural, because there are silence gaps between the central vowels and adjacent steady-states.

The question in the present experiments is how important are such cues as adjacent steady-state sounds, formant transitions or both of them with regard to overshoot or extrapolation of the center vowel. To solve this question, various stimulus conditions are used in a center vowel matching and identification test to measure the amount of perceptual vowel boundary shifts. The central vowels are 5-formant type synthesized vowels and the surrounding steady-state and transition are either single-formant type sounds or 5-formant type vowels to observe whether extrapolation is induced more by speech-like sounds.

II. Experiment 1

The first experiment studies how single-formant or 5-formant steady-states and transitions as pre- and post-anchors can affect central vowel perception by using a vowel identification test.

A. Stimuli:

The stimuli were synthesized under four conditions:

1) 5-formant stationary vowels of 120-ms duration, situated on the Japanese /u/-/e/-/a/ continuum by varying F1 (See Table 1), presented in isolation (called Ref.)
2) stationary vowels preceded and followed by a connecting 30-ms transition (called Λ)
3) stationary vowels preceded and followed by a 120-ms steady state, separated by a 30-ms gap (called Π)
4) stationary vowels preceded and followed by a 120-ms steady state, connected with a 30-ms transition (called Ω)

All of the stimuli were synthesized with a stationary pitch of 140 Hz at a sampling frequency of 20 kHz. Figure 1 shows schematic illustrations of four possible contour lines of an F1, or actually of any single formant. The steady-state and the transition are a single-formant sound (Stimulus set I) or a 5-formant type vowel (Stimulus set II).

Stimulus set I (single-formant anchor) is used to observe how single-formant surrounding sounds influence central vowel perception. The surrounding steady-state sounds in Stimulus set I were synthesized under the following way. The Klatt cascade formant synthesizer with 5 formants was used for synthesis and the bandwidth of the F2 - F5 was substantially expanded. Thus, the F2 - F5 frequencies were the same as for the central vowel condition and the F2 - F5 bandwidths were 10,000 Hz, the F1 frequency was 2.73 Bark (281 Hz), and the bandwidth was 50 Hz. Transition portions of Stimulus set I were synthesized under the same condition as the steady-state, but only the F1 frequency was

Table 1. Formant frequency and bandwidth for central vowel synthesis with in Bark and in Hz.

Stimulus Number	Formant	Center Frequency Bark[12]	Hz	Band Width Hz
0		2.93	301	
1		3.13	322	
2		3.33	343	
3		3.53	365	
4		3.73	387	
5		3.93	409	
6		4.13	431	
7		4.33	454	
8		4.53	477	
9	F1	4.73	500	50
10		4.93	523	
11		5.13	546	
12		5.33	569	
13		5.53	593	
14		5.73	618	
15		5.93	643	
16		6.13	668	
17		6.33	693	
18		6.53	719	
19		6.73	745	
	F2	11.20	1500	150
	F3	14.50	2500	250
	F4	16.50	3500	350
	F5	17.90	4500	450

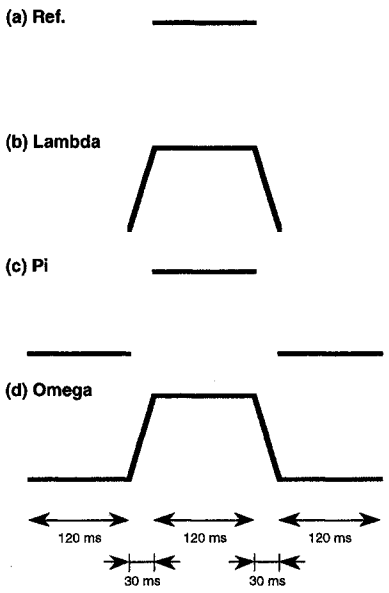

Figure 1. Schematic figure for single-formant contour, (a) vowel presented in isolation (called Ref.), (b) [transition + vowel + transition] (called L), (c) [steady-state + gap + vowel + gap + steady-state] (called P) and (d) [steady-state + transition + vowel + transition + steady-state] (called Ω).

swept from 2.73 Bark (281 Hz) to the central vowel F1 frequency and back again. Thus, the transition rate was 0.7 - 18.1 Hz/ms.

Stimulus set II (five-formant anchors) is used to observe how speech-like surrounding sounds affect central vowel perception. Both central vowel and gap are the same as in Stimulus set I. F1 of the steady-state surrounding sound and of the transition sound are also the same as in Stimulus set I, but the F2 - F5 are the same as in the central vowel condition.

Hereafter, single-formant type stimuli are called Λ, Π and Ω and 5 formant type stimuli are called Λ5, Π5 and Ω5.

B. Subjects:

The subjects were 10 graduate course students at Japan Advanced Institute of Science and Technology. All subjects were native speakers of Japanese with no known hearing impairment.

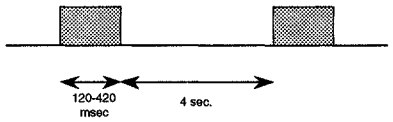

Figure 2. Paradigm for the identification experiment.

C. Procedure:

The stimuli in each type were randomized and recorded on a DAT player (SONY DTC-57ES) at 4-s intervals as shown in Fig. 2. There was a 1000 Hz, 50-ms pure tone after every 10 trials and 8-s pause after every 100 trials. Each stimulus was repeated 20 times in one session. Thus, there were 400 (20 F1 × 20 times) stimuli presented in one session. The experimental DAT tapes were reproduced on a DAT player (SONY DTC-57ES) and presented through STAX SR Λpro headphones in a sound proof room (22.7 dB (A)).

The subjects were required to identify the central vowel in each stimulus as either the vowel /u/, /e/, or /a/. Each subject listened to 3 sessions in 3 months. The results were used to determine the phoneme boundary between /u/ and /e/, or /e/ and /a/, by fitting the percentage of /u/ or /a/ judgments to an integrated Gaussian function. The point at which the /u/ judgments fell below 50% or at which the /a/ judgments exceeded 50% was regarded as the boundary.

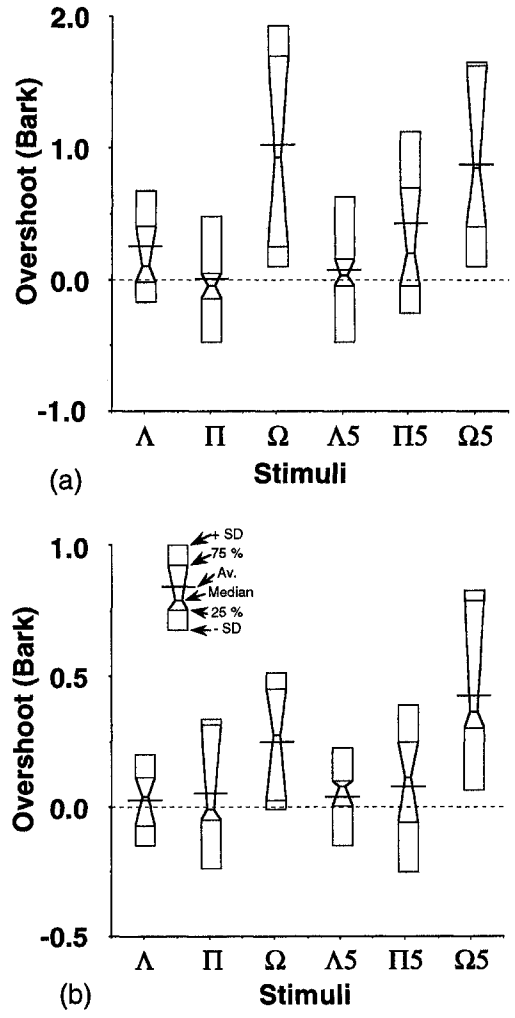

Figure 3. Boundary shifts for the three indicated stimulus types with various pre- and post-anchors between (a) /u/ and /e/, and (b) /e/ and /a/ relative to the boundaries for the Ref. type stimuli.

Table 2. F-test results, (a) between stimulus types and (b) between stimulus sets. * indicates values larger than F(1,18; 0.05) = 4.41.

(a)	Λ vs. Π	Π vs. Ω	Λ vs. Ω	Λ5 vs. Π5	Π5 vs. Ω5	Λ5 vs. Ω5
/u/-/e/	1.50	9.49*	5.58*	1.63	1.79	7.05
/e/-/a/	0.04	2.92	4.48*	0.04	7.28*	10.51*

(b)	Λ vs. Λ5	Π vs. Π5	Ω vs. Ω5
/u/-/e/	0.66	2.55	0.12
/e/-/a/	0.04	0.02	1.39

D. Results:

Since the phoneme boundaries of each stimulus type were consistent in 3 sessions during 3 months, they were taken together to determine averages of phoneme boundaries for each stimulus type. Normal phoneme boundaries for /u/-/e/ and /e/-/a/ were determined by using the response for the Ref. type stimuli for each subject. The average normal phoneme boundary is 4.92 Bark (SD = 0.47 Bark) in the /u/-/e/ condition and 5.87 Bark (SD = 0.26 Bark) in the /e/-/a/ condition. The subjects had no difficulty to identify the vowels and the results indicate that the subjects perceived the vowels very categorically. However, there were some individual differences[10] as the standard deviation shows. The amount of overshoot is calculated by subtracting from the phoneme boundary for the Λ, Π or Ω type stimuli the normal (Ref. type) phoneme boundary in /u/-/e/ or /e/-/a/ condition for every subject.

Figure 3 shows the experimental results. The vertical axis indicates boundary shifts with pre- and post-anchors between (a) /u/ and /e/ or (b) /e/ and /a/ on a Bark scale. Positive value occur when the other stimulus boundaries are higher than the Ref. type boundary. Thus, positive values indicate overshoot or extrapolation between steady-states and/or transitions and the central vowel. Table 2 shows F-test results between stimulus type pairs.

The experimental results in both /u/-/e/ and /e/-/a/ boundaries are similar, which shows that:

Central vowel extrapolation occurs with Ω type stimuli in both single- and 5-formant conditions, whereas some averaging effects are observed with Λ and Π type stimuli for some of the subjects. The overall order of the amount of overshoot is Ω > Λ > Ref. > Π in the single-formant condition and Ω5 > Π5 > Λ5 > Ref. in the 5-formant condition.

However, some aspects were different between /u/-/e/ and /e/-/a/ boundaries:

The amount of overshoot with a 5-formant steady-state for the Π type stimuli of the /u/-/e/ boundary is larger than with a single formant steady-state, and for the Ω type stimuli of the /u/-/e/ boundary it is a little bit larger than with a single formant Ω type stimulus.

III. Experiment 2

To detect a perceptual equivalence point between the four stimulus types, the second experiment was performed by using a matching test.

A. Stimuli and Subjects:

Stimuli used for experiment 2 are the same as those for experiment 1. The subjects were 10 graduate course students. Nine of the ten subjects that served in experiment 1 also served in experiment 2.

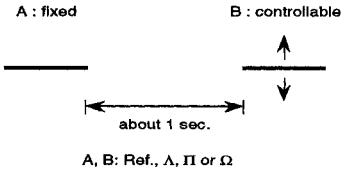

Figure 4. Paradigm for the matching experiment.

B. Procedure:

The four types of stimuli were paired against each other as shown in Fig. 4. Sound A is fixed to be stimulus number 11, that is, the F1 of the central vowel is 5.13 Bark (546 Hz), whereas sound B is varied under control of the subject. Subjects have to tune sound B in such a way that it is as similar as possible to sound A at the central vowel position. Each of the A and B stimulus pairs was presented 8 times in one sesstion randomly through STAX SR Λpro headphones in a sound proof room (22.7 dB (A)). The number of stimuli was 128 (4 types × 4 types (16 pairs) × 8 times)

C. Results:

The means (M) and the standard deviations (SD) of the difference between the perceptual equivalence point and the physical equivalence point, i.e., the difference between the F1 in Bark of the tuned sound B stimulus and the F1 in Bark of fixed sound A (stimulus 11), are calculated (See Table 1) for each of the A and B stimulus pairs that were presented 8 times. The ranges of M and SD in the 4 carrier conditions, Ref vs. Ref, Λ vs. Λ, Π vs. Π and Ω vs. Ω, are less than 0.1 Bark and 0.15 Bark, respectively. The results of other stimulus pairs show no effects or averaging, not even with the Ω type stimuli. Table 3 shows the amount of overshoot in the Ref vs. Ω condition for every subject and Table 4 shows responses for the subject KA, which is a typical case. The tables indicate that the results of the identification test and the matching test are different very much. In the next section, these results are discussed.

IV. Discussions

A. Relations between Ref., Λ, Π and Ω

The results with Λ and Λ5 type stimuli show very small amounts of overshoot. This indicates that adjacent transition portions affect the perception of central vowels, whereas the amount of overshoot is very small when the anchors consist of transitions only. On the other hand, averaging effects with Λ and Λ5 type

Table 3. Matching and identification test results, the amount of overshoot in the Ref vs. Ω condition for the 9 subjects that participated in both experiments.

Effect	Matching test	Identification test
overshoot (0.2 Bark <)		IG,KI,KU,KA,NO,OG
weak overshoot (0.1 Bark <)	AS	AS,YO
nothing	KI,KA,YO,OG	KO
weak averaging (< -0.1 Bark)	IG,KU	
averaging (< -0.2 Bark)	NO,KO	

Table 4. Matching experiment results for the subject KA. The left column indicates sound A and the top row indicates sound B. The upper value in an element shows a central vowel F1 difference between sound A and B and the lower value shows a standard deviation. If the difference is positive, sound A is perceived with more overshoot than sound B.

		B			
	Average SD (Bark)	Ref.	Λ	Π	Ω
A	Ref.	0.04 0.10	0.01 0.14	0.04 0.14	0.09 0.17
	Λ	0.07 0.10	-0.01 0.14	0.03 0.12	-0.05 0.14
	Π	0.08 0.12	0.08 0.14	-0.02 0.14	0.06 0.16
	Ω	0.07 0.13	-0.07 0.18	-0.01 0.10	-0.04 0.14

stimuli were observed for some of the subjects. This supports claims of Huang[3] and Pols and van Son[4] that some averaging is observed in the perception of F1 time-varying trajectories.

Averaging effects with the Π type stimuli were also observed for a considerable number of subjects. Previous investigations using [single-formant pre-anchor + gap + vowel] type stimuli[8][10][11] showed an overshoot effect. Thus, it was expected that an overshoot effect would appear for the present Π stimulus condition as well. However, less overshoot effects were observed with the Π type stimuli, possibly because (1) the Π type stimuli are artificial and strange, or because (2) pre- and post-steady-states without transitions pull down the central vowel, although the central vowel contrasts with the pre- steady-state.

The amount of overshoot for the Ω type stimuli is the largest and is larger than the sum of that for the Λ and the Π type stimuli. This suggests that each adjacent steady-state or transition does not affect perception of the central vowel very much, and that connecting the steady-state pre-anchor, center vowel and steady-state post-anchor and filling gap with transition induces extrapolation most, because the Ω type stimuli sound the most natural, and subjects could perceive the Ω type stimuli as a natural stream and the Π type stimuli as a connected artificial sound.

B. Single formant vs. 5 formant

Although the value of F-test between Π and Π5 of /u/-/e/ boundary in Table 2(b) is 2.55 which is smaller than 4.41 with F(1,18;0.05), this is larger than other pairs and the mean values are far apart between Π and Π5 of /u/-/e/ boundary (See Fig. 3(a)). This indicates that the amount of overshoot is also related to the "vowelness" of pre- and post-anchors, because the Π type stimuli are artificial and strange. Still there may be a phonemic interaction in the 5-formant type condition. These findings support the idea that an overshoot becomes more intense with increased stimulus complexity[8].

The reason that overshoot is observed in the /u/-/e/ boundary, and not in the /e/-/a/ may be as follows. Akagi[10] reported that: the amount of overshoot becomes maximum when the F1 difference between the preceding vowel anchor and the perceived one is about 2 Bark, whereas the amount of overshoot becomes very small when the difference becomes larger than 3 Bark. Here, the perceptual mean F1 frequency of the /u/-/e/ boundary is 4.92 Bark, that of the /e/-/a/ boundary is 5.87 Bark and that of the pre- and post-anchor frequency is 2.73 Bark. The differences are 2.2 Bark in the /u/-/e/ boundary and 3.2 Bark in the /e/-/a/ boundary. Thus, the overshoot is observed in the /u/-/e/ boundary of the Π5 type stimuli. Additionally, the amount of overshoot using Ω type stimuli in the /u/-/e/ boundary is twice as large as that in the /e/-/a/ boundary.

C. Identification test vs. Matching test

There is much difference between the results of the identification test and the matching test, even for the same subjects mainly (See Table 3). Overshoot effects are observed in the identification test, whereas the averaging effects are observed in the matching test. It is expected that in the identification test, the subjects were forced to perceive stimuli as a phoneme (vowel), whereas in the matching test, they may not perceive the sounds phonetically but analytically. These findings suggest that sounds have to be perceived as phonemes to observe overshoot. So, most probably, overshoot will only be observed in natural speech (possibly to compensate for articulatory undershoot), however, only under some contextual conditions. It results only from a higher-order central processing mechanism.

V Conclusion

The vowel identification and matching experiments were performed to examine how preceding and following anchor signals affect central vowel perception. The results of both the experiment suggest that:

(1) central vowel overshoot occurs with Ω-type stimuli, in both single- and 5-formant conditions, whereas averaging effects are observed with Λ- and Π-type stimuli for some of the subjects. The overall order of the amount of overshoot is Ω > Π > Λ > Ref. in the 5-formant condition. The most natural-sounding Ω-type stimuli showed the largest overshoot,

(2) especially for the Π-type stimuli, the amount of overshoot with a 5-formant steady-state is larger than with a single-formant steady-state. This might be an indication that also the 'vowelness' of the pre- and post-anchors contributes to the amount of overshoot, and

(3) it is significant for overshoot that central vowels are perceived as phonemes.

Acknowledgment

The authors would like to thank Mr. Ryuji Nakagawa of Japan Advanced Institute of Science and Technology for conducting the psychoacoustic experiments in part reported in Section II. The forst author spent some time in Amsterdam on a grant from the JAIST foundation to prepare the experiments and to do some pilots, whereas some equipment used in this research was supported in part by grant no. 931001 from the Tateisi Science and Technology Foundation.

References

[1] Lindblom, B. E. F. and Studdert-Kennedy, M., "On the role of formant transition in vowel recognition", J. Acoust. Soc. Am., 42, 830-843, 1967.

[2] Kuwabara, H and Sakai, H., "An experiment on the phoneme boundary locations in the dynamic perception of synthetic vowels", J. Acoust. Soc. Jpn., 31 (1), 18-23, 1975. (In Japanese with English abstract).

[3] Huang, C. B., "Modeling Human vowel identification using aspects of formant trajectory and context", Speech Perception, Production and Linguistic Structure (Tohkura, Y., Vatikiotis-Bateson, E., and Sagisaka, Y. Eds.), Ohmsha Tokyo and IOS Press Amsterdam, 43-61, 1992.

[4] Pols, L. C. W. and van Son, R. J. J. H., "Acoustics and perception of dynamic vowel segments", Speech Communication 13, 135-147, 1993.

[5] Suzuki, H., "Mutually complementary effect between amount and rate of formant transition in perception of vowels, semivowels and voiced stops and a possible mechanism for their identification", J. Acoust. Soc. Jpn., 30 (3), 169-180, 1974. (In Japanese with English abstract).

[6] Strange, W., Jenkins, J., and Johnson, T. L., "Dynamic specification of coarticulated vowels", J. Acoust. Soc. Am., 74, 695-705, 1983.

[7] Furui, S., "On the role of spectral transition for speech perception", J. Acoust. Soc. Am., 80, 1016-1025, 1986.

[8] Shigeno, S. and Fujisaki, H., "Effect of a preceding anchor upon the categorical judgment of speech and non-speech stimuli", Japanese Psychological Research, 21, 165-173, 1979.

[9] Shigeno, S., "Assimilation and contrast in the phonetic perception of vowels", J. Acoust. Soc. Am., 90, 103-111, 1991.

[10] Akagi, M., "Psychoacoustic evidence for contextual effect models", Speech Perception, Production and Linguistic Structure (Tohkura, Y., Vatikiotis-Bateson, E., and Sagisaka, Y. Eds.), Ohmsha Tokyo and IOS Press Amsterdam, 62-78, 1992.

[11] Akagi, M., "Modeling of contextual effects based on spectral peak interaction", J. Acoust. Soc. Am., 93, 1076-1086, 1993.

[12] Zwicker, E. and Terhaldt, E., "Analytical expressions for critical-band rate and critical band-width as a function of frequency", J. Acoust. Soc. Am., 65, 1523-1525, 1980.