An objective critical distance measure based on the relative level of spectral valley

T V Ananthapadmanabha[1], *A G Ramakrishnan*[2], *Shubham Sharma*[3]

[1]Voice and Speech Systems, Malleswaram, Bangalore, India
[2]Department of Electrical Engineering, Indian Institute of Science, Bangalore, India
[3]Department of Electrical Communication Engineering, Indian Institute of Science, Bangalore, India
tva.blr@gmail.com, ramkiag@ee.iisc.ernet.in, shubham@ece.iisc.ernet.in

Abstract

Spectral integration is a subjective phenomenon in which a vowel with two formants, spaced below a critical distance, is perceived to be of the same phonetic quality as that of a vowel with a single formant. It is tedious to conduct perceptual tests to determine the critical distance for various experimental conditions. To alleviate this difficulty, we propose an objective critical distance (OCD) that can be determined from the spectral envelope of a speech signal. OCD is defined as that spacing between the adjacent formants when the level of the spectral valley between them reaches the mean spectral value. The measured OCD lies in the same range of 3 to 3.5 Bark as the subjective critical distance for similar experimental conditions giving credibility to the definition. However, it is noted that OCD for front vowels is significantly different from that for the back vowels.
Index Terms: vowel perception, spectral analysis, critical distance, spectral integration, spectral valley

1. Introduction

Delattre *et al.* of Haskins Labs reported in 1952 an interesting experimental study [1] of vowel quality. They showed that the perceptual quality of a synthesized back vowel with two formants matches well with that of a synthesized vowel with an appropriately chosen single formant. This equivalence of phonetic quality of a two-formant vowel to that of a single formant vowel is referred to in the literature as "spectral integration". Surprisingly, this integration did not seem to emerge in the case of front vowels, except for vowel 'i' as an extreme case. Chistovich *et al.* [2] conducted subjective experiments to derive the spacing between formants, below which the spectral integration occurs. This spacing is referred to as the critical distance and is denoted by ΔZ_c. In their experiments, they kept F_2 fixed and varied the spacing between F_1 and F_2. Such a two-formant stimulus was compared with a single-formant stimulus for equivalence in vowel quality. They found that spectral integration occurs when $3.1 < \Delta Z_c < 4.0$ Bark for an F_2 value of 1.8 kHz and $3.3 < \Delta Z_c < 4.3$ Bark for an F_2 value of 1.4 kHz. Although ΔZ_c is F_2 dependent, has a wide range of 3.1 to 4.3 Bark and for a synthetic vowel /u/, spectral integration is known to occur for a formant separation of 3.94 Bark [1], this perceptual phenomenon is generally known as the 3-Bark rule in the literature.

Chistovich *et al.* [2] argued that the frequency of the single-formant equivalent corresponds to the center of gravity of the spectrum. Chistovich and Lublinskaya [3] showed that spectral integration occurs even when there is a large change in formant levels when the spacing between the formants is less than the critical distance in the range of $3 - 3.5$ Bark. Experiments on matching a four-formant vowel to a two-formant vowel [4] support the so called 3-Bark rule hypothesis. Some researchers consider spectral integration to be a general psycho-acoustic phenomenon, not necessarily restricted to vowels, based on perceptual tests conducted on two-tone complex signals [5] and sinusoids replacing formants [6, 7]. Spectral integration has also been reported in the literature with respect to syllable initial stops [8] and glides [9]. An exception to spectral integration has been demonstrated for a specific back vowel of Chinese language [10].

Motivation for the present study: Generally, the subjective critical distance has been determined using two-formant unrounded synthetic vowels resembling the quality of /ɑ/ or /æ/. It is not known if the same subjective critical distance is also valid for front vowels and for multiple-formant natural vowels with different bandwidths, formant levels, $F0$ etc. It is extremely time consuming to conduct perceptual tests to determine the critical distance for all the experimental conditions. With this in view, we propose an objective critical distance (OCD) that may be measured for a given spectral envelope and investigate if such an OCD serves a purpose same as that of the subjective critical distance.

About this work: In Section 2, OCD is defined in terms of the level of spectral valley between two formants relative to the mean spectral value. We replicate the experimental conditions for which perceptually determined subjective critical distance (SCD) is available. Also, we study the influence of formant spacing, higher formants, formant levels and fundamental frequency on the measured OCD. In Section 3, the measured OCD is compared with the SCD published in the literature.

2. The objective critical distance

2.1. Rationale

The rationale for arriving at a definition for OCD relies on the SCD as follows: (a) SCD depends the formant spacing. We note that the level of the spectral valley (LSV) between the formants is determined by the spacing between the formants. (b) SCD is independent of the formant levels. We show later (in Sec. 2.5.1) that LSV varies only by a couple of dB even when the formant levels vary over a wide range of about 12 to 20 dB. (c) SCD is about 3 Bark. We have to find a suitable value to act as a threshold such that LSV is higher (or lower) than the threshold when the spacing between the formants is less (or greater) than 3 Bark. (d) In subjective experiments, a two-formant vowel stimulus is compared with a single-formant stimulus keeping the energy of the two stimuli nearly the same. The energy of a signal determines the *mean spectral value* for a given bandwidth and is used as the threshold. We now postulate a formal definition of OCD.

2.2. Definition of OCD

For a given magnitude squared spectrum, we define the relative level of spectral valley (RLSV) between the formants F_1 and F_2, denoted by V_{12}, as the ratio of the mean spectral value to the level of the valley between F_1 and F_2:

$$V_{12} = S_m / S_{v_{12}} \quad (1)$$

where S_m is the mean value of the spectrum and $S_{v_{12}}$ is the level of the spectral valley. RLSV is expressed in dB. As the spacing between the formants is varied, RLSV also varies. We define OCD, denoted by ΔF_{12}, as that value of the separation between F_1 and F_2 when V_{12} becomes 0 dB. That is,

$$\Delta F_{12} = F_2 - F_1, \quad (2)$$

when $S_{v_{12}}$ is the same as S_m. The above definition can be extended for the case of any two adjacent formants, such as F_2 and F_3. So,

$$\Delta F_{23} = F_3 - F_2, \quad (3)$$

when the level of the spectral valley $S_{v_{23}}$ between F_2 and F_3 is the same as S_m.

2.3. The influence of formant spacing on RLSV in a two-formant vowel

A two-formant vowel is synthesized using a second order digital resonator with F_2 and B_2 kept constant at 1400 and 200 Hz, respectively. The frequency of the first formant F_1 is varied from 650 to 950 Hz in steps of 50 Hz and its bandwidth is kept constant at 100 Hz. The reason for this choice of parameters is to replicate the experiment conducted by Chistovich *et al* [2]. The magnitude squared spectrum of the impulse response of the cascaded two-formant model is computed. Figures 1(a) to 1(c) show the log-magnitude spectra for three values of F_1, namely, 650, 750 and 850 Hz, respectively. For F_1 = 750 Hz (case b) V_{12} is close to 0 dB and the formant spacing is about 3 Bark, which as per the proposed definition corresponds to OCD. Figure 2 shows V_{12}, as a function of formant separation in Bark. It is not surprising that V_{12} increases with the increase of formant spacing. However, it is worth noting that V_{12} is nearly 0 dB when the spacing between the two formants is about 3 Bark.

Although both the formant peaks are clearly resolved in the log-magnitude spectra for all the three conditions, the spectral integration resulting in a single formant perception occurs only for cases (b) and (c) where the formant spacing is equal to or less than the OCD[1].

2.4. The influence of higher formants on OCD

We consider a four-formant vowel with the formant frequencies at 500, 1500, 2500 and 3500 Hz, corresponding to those of a uniform tube. As will be discussed in Sec. 2.5.1, the choice of bandwidth does not significantly influence ΔF_{12}. Hence, for simplicity, bandwidths for all the four formants are kept fixed at 100 Hz. Any other choice for bandwidths could have been made without significantly affecting the results. Also, it is not necessary that the bandwidths of all the formants be equal. In order to control the spacing between F_1 and F_2, F_1 is increased and simultaneously F_2 is decreased in steps of 25 Hz. It is noted that there exists a condition where the measured V_{12} is less than or nearly equal to 0 dB with OCD to be 3.6 Bark, even for the case of a four formant vowel. The estimate of OCD matches well with the reported results based on subjective experiments [3, 5].

[1]The code used and wav files generated for the experiments of Sec. 2 are available at *http://mile.ee.iisc.ernet.in/mile/download.html*

2.5. The influence of formant levels and F0 on OCD

For synthesis, a four-formant model with $F_3 = 2500$ Hz, $F_4 = 3500$ Hz and a sampling frequency of 8000 Hz has been used. In order to replicate a previous study [3], two different cases of formant spacing are considered:

Case (a), where F_1, F_2 spacing is 2.5 Bark with $F_1 = 400$ Hz and $F_2 = 700$ Hz.

Case (b), where F_1, F_2 spacing is 4.65 Bark with $F_1 = 600$ Hz and $F_2 = 1300$ Hz.

2.5.1. The influence of formant levels via bandwidth changes

Since we are using a cascaded formant model, we can obtain different formant levels by controlling the bandwidths. The bandwidths (B_1 and B_2) of the first two formants are varied over a wide range to obtain different spectral levels in dB, denoted as L_1 and L_2, respectively. B_3 and B_4 are fixed at 100 Hz. The magnitude squared spectrum of the impulse response of the cascaded four-formant model is computed. Figure 3 shows the measured V_{12} as a function of the difference in formant levels in dB. For case (a), where F_1, F_2 spacing is less than 3-Bark, it is seen that V_{12} in dB is consistently negative for $(L_1 - L_2)$ range of about ± 6 dB. For case (b), where F_1, F_2 spacing is greater than 3-Bark, it is consistently positive for $(L_1 - L_2)$ range of about ± 10 dB. Though $(L_1 - L_2)$ varies over a wide range of 12 to 20 dB, V_{12} varies only by about 2 dB. Thus a large change in bandwidths or in the formant levels does not significantly influence V_{12} and hence the estimated OCD.

2.5.2. The influence of fundamental frequency on OCD

For each case of formant spacing, initially the impulse response is computed to obtain a reference value of V_{12} denoted by $V_{12_{ref}}$. The response of the 4-formant filter (cases a and b) is computed for a source input of a periodic sequence of impulses for various choices of pitch period (or the fundamental frequency $F0$). The bandwidths of all the formants are kept constant at 100 Hz. Smoothed spectral envelope is obtained using linear prediction of order 8 to measure RLSV as a function of $F0$ denoted by $V_{12_{F0}}$. The results are shown in Table 1. The difference $(V_{12_{ref}} - V_{12_{F0}})$ is within ± 1 dB for the formant spacing of case (a) whereas, it is consistently positive and above $+2$ dB for the formant spacing of case (b). This demonstrates that $F0$ does not have a significant influence on the estimated OCD.

Table 1: *Fundamental frequency $F0$ does not significantly influence the estimation of OCD. The Table shows the deviation in the estimated V_{12} (in dB) for a periodic vowel from that of the reference impulse response.*

F0 (Hz)	**100**	**125**	**150**	**175**	**200**	**225**	**250**
$\mathbf{F_2 - F_1 < 3}$ **Bark**	-0.30	-0.49	0.09	-0.48	-0.78	0.99	-0.51
$\mathbf{F_2 - F_1 > 3}$ **Bark**	2.23	2.44	2.29	2.79	2.60	2.45	2.32

2.6. Objective critical distance for natural vowels

In order to study RLSV for different natural vowels, we use the mean values of the first three formants of vowels of American English published by Peterson and Barney [11, 12]. We have

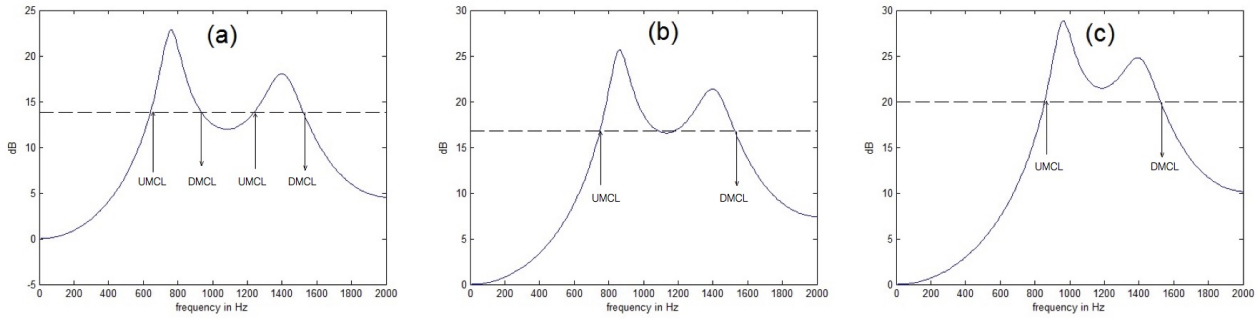

Figure 1: *Log-magnitude spectrum of a synthetic two-formant vowel for three values of formant spacing in Bark, namely, (a) 3.96 (b) 3.2 (c) 2.55. The level of the spectral valley relative to the mean spectral value (dashed line) for the three cases are -1.9, -0.3 and 1.5 dB, respectively. The frequencies, where the log spectrum crosses the mean spectral value, are shown by vertical lines.*

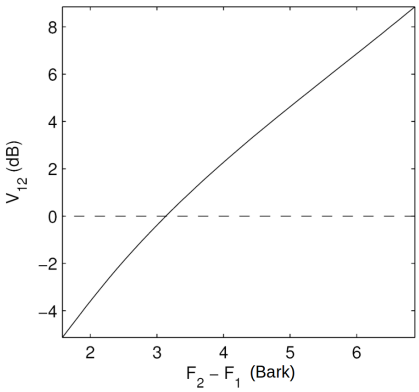

Figure 2: *The influence of formant spacing on the relative level of the spectral valley (in dB).*

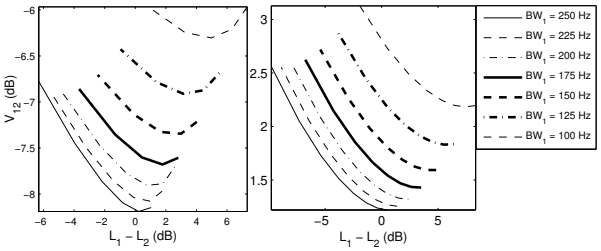

Figure 3: *The influence of formant levels on the relative level of the spectral valley. For each value of BW_1, BW_2 is varied over a wide range. (a) (F_1, F_2) separation $<$ 3 Bark and (b) (F_1, F_2) separation $>$ 3 Bark.*

synthesized four-formant vowels with F_4 as 3500 Hz for adult male and 4200 Hz for the adult female speaker. Bandwidth of 100 Hz is used for all the formants. Sampling frequencies of 8000 Hz and 10000 Hz are used for the adult male and female speakers, respectively. For back vowels, we begin with the mean formant values of a vowel and then vary the formant spacing between F_1 and F_2 by increasing (decreasing) F_1 while simultaneously decreasing (increasing) F_2 to the same extent. For front vowels, the spacing between F_2 and F_3 is varied in a similar manner. The log-magnitude spectrum is computed from the impulse response of the four-formant vowel. The measured OCD values in Bark are listed in Table 2 for nine vowels. For the experiment on back vowels, OCD varies in the range of 3.6 to 4.9 Bark with the varying separation between F_1 and F_2. For the experiment on front vowels, OCD varies in the range of 1 to 1.8 Bark with the varying separation between F_2 and F_3. The values of the OCD are comparable for male and female speakers. A strong vowel dependency is seen.

3. A comparison of OCD with SCD

Influence of formant spacing: In Sec. 2.3, formant data similar to those used by Chistovich *et al.* [2] have been used. In their study, two parallel bandpass filters were used, with F_2 kept fixed while varying F_1. The level of F_2 was lower than that of F_1 and hence we have used B_2 = 200 Hz and B_1 = 100 Hz. They observed that spectral integration takes place when the spacing between the formants is less than an average critical distance of about 3 Bark. The measured OCD based on spectral valley criterion proposed in this paper, ΔF_{12}, for similar experimental conditions is about 3.2 Bark, which matches well with the reported average ΔZ_c.

Influence of formant levels: In Sec. 2.5.1, we have used formant data and levels similar to those used by Chistovich and Lublinskaya [3]. In their study, two parallel bandpass filters were used and the gains (A_1, A_2) of the formants were altered. For a given A_2/A_1, the resonant frequency, F* of a single formant stimulus is varied so as to match the subjective vowel quality to that of the two-formant stimulus. The matched F* lies between F_1 and F_2 when the spacing between the formants is less than or equal to the critical distance over a wide range of A_2/A_1. Thus, spectral integration is shown to be applicable over a wide range of formant levels. When the formant spacing is greater than ΔZ_c, the responses of the two subjects participating in the experiments were inconsistent. In Sec. 2.5.1, we have noted that V_{12} in dB is consistently negative over a wide range of formant levels when the spacing between formants is less than or equal to 3 Bark. When the spacing is greater than 3 Bark, V_{12} in dB is consistently positive. The objective experiment reported in Sec. 2.5.1 gives results similar to those reported in the above subjective study [3]. It would be interesting to study if the RLSV is insensitive also to (F_2, F_3) separation.

Effect of (F_1, F_2) separation in natural back-vowels: The estimated OCD, ΔF_{12}, for rounded back vowels lies in the range of 3.9 to 4.9 Bark, which is much higher than the usually reported subjective critical distance of 3 to 3.5 Bark. The high value observed for the OCD for rounded back vowels gets sup-

port from other studies: (i) In the original seminal work of Delattre *et al.* [1], it has been shown that spectral integration does take place for vowel /u/ with F_1 = 250 Hz and F_2 = 700 Hz, having a separation of 3.94 Bark, which is greater than the reported subjective critical distance. (ii) The reported formant spacing between F_1 and F_2 for back vowels [11], for both male and female speakers, lies in the range of 3.8 to 5.0 Bark [13]. (iii) In a subjective experiment on a series of rounded back vowel stimuli [14], (F_1, F_2) separation in the range of 4.2 to 4.7 Bark has been used. These evidences, along with the high value of the measured OCD, motivates one to inquire if a constant critical distance of 3 or 3.5 Bark can be universally applied for all the experimental conditions.

Effect of ($\mathbf{F_2}$, $\mathbf{F_3}$) separation in natural front-vowels: To our knowledge, there is no reported study on determining the subjective critical distance for front vowels based on (F_2, F_3) separation, though there have been some studies on the equivalent F_2 determination of vowels [15, 16]. However, in an experiment on front vowel series [14], 'hid' versus 'head', spectral integration is shown to occur, for (F_2, F_3) separation in the range of 1.2 to 1.8 Bark, which compares well with the OCD derived for front vowels, shown in Table 2.

Table 2: *The estimated OCD values (in Bark) for natural vowels.* $\exists$ *denotes a uniform tube.*

Back Vowel $\mathbf{F_2}$ - $\mathbf{F_1}$ spacing	/ɑ/	/ɔ/	/ʊ/	/u/	/ʌ/
ΔF_{12} (male)	3.95	4.57	4.58	4.56	4.01
ΔF_{12} (female)	4.33	4.89	4.86	4.93	4.26
Front Vowel $\mathbf{F_3}$ - $\mathbf{F_2}$ spacing	/i/	/ɪ/	/ɛ/	/æ/	$\exists$
ΔF_{23} (male)	1.05	1.50	1.66	1.79	1.82
ΔF_{23} (female)	1.08	1.32	1.46	1.67	1.82

4. Conclusion

A term called objective critical distance has been defined as that spacing when the level of the spectral valley between two formants is equal to the mean spectral value. We have shown that the behaviour of the OCD is similar to the subjectively derived critical distance for similar experimental conditions.

We have shown that the level difference between the first two spectral valleys is useful for determining the front/back phonetic feature of vowel sounds [17]. This speaker independent method doesn't require an explicit extraction of formant frequencies.

We define the upward mean-level crossing, UMLC, as the frequency where the log-spectrum crosses the mean spectral value, from a lower value to a higher value. Similarly, we define the downward mean-level crossing, DMLC. These mean-level crossings (MLCs) are shown by arrows in Figure 1 for a two-formant vowel. Here, when the spacing between formants is greater than OCD, (i.e., when two formants are perceived) there are two pairs of UMLC-DMLC. When formant spacing is equal to or less than the OCD, (i.e., when only one formant is perceived) there is only one pair of UMLC-DMLC. Based on this observation, we conjecture (a) one formant is perceived for every pair of UMLC-DMLC (b) the detailed spectral shape between UMLC-DMLC (spectrum above mean spectral value) is unimportant since the perceived single formant is independent of the formant levels and (c) formants below the mean spectral value produce no MLCs and hence may not be perceived at all and may only play a role in determining the mean spectral value and the spectral tilt. If such a conjecture were to be established by perceptual experiments, it implies that an auditory spectral envelope is distinct from that of an acoustic spectral envelope with implications both for speech perception as well as automatic speech recognition.

5. References

[1] P. Delattre, A. Liberman, F. Cooper and L. Gerstman, "An experimental study of the acoustic determinants of vowel color; observation on one- and two-formant vowels synthesized from spectrographic patterns," in *Journal of Acoustical Society of America*, vol. 8, no. 3, pp. 195–210, 1952.

[2] L. Chistovich, R. Sheikin and V. Lublinskaya, "Centres of gravity and spectral peaks as the determinants of vowel quality," in *Frontiers of Speech Communication Research*, edited by B. Lindblom, S. E. G. Ohman and G. Fant, pp. 143–157, Academic Press, 1979.

[3] L. A. Chistovich, and V. V. Lublinskaya, "The 'center of gravity' effect in vowel spectra and critical distance between the formants: psychoacoustical study of the perception of vowel-like stimuli," in *Hearing Research, Elsevier*, vol 1, no. 3, pp. 185–195, 1979.

[4] R. Carlson, G. Fant and B. Granstrom, "Two-formant models, pitch, and vowel perception," in *Auditory analysis and perception of speech*, edited by G. Fant and M.A.A. Tatham, New York: Academic Press, pp. 55–82, 1975.

[5] Q. Xu, E. Jacewicz, L. L. Feth and A. K. Krishnamurthy, "Bandwidth of spectral resolution for two-formant synthetic vowels and two-tone complex signals," in *Journal of Acoustical Society of America*, vol. 115, no. 4, pp. 1653–1664, 2004.

[6] R. A. Fox, E. Jacewicz and C.-Y Chang, "Vowel perception with virtual formants," in *XVI International Congress of Phonetic Sciences*, pp. 689–692, 2007.

[7] R. A. Fox, E. Jacewicz and C.-Y Chang, "Auditory spectral integration in the perception of static vowels," in *Journal of Speech Language Hearing Research*, vol. 54, no. 6, pp. 1667–1681, 2011.

[8] R. A. Fox, E. Jacewicz and L. L. Feth, "Spectral integration of dynamic cues in the perception of syllable-initial stops," in *Phonetica*, vol. 65, no. 1-2, pp. 19–44, 2008.

[9] R. A. Fox, E. Jacewicz and C.Y Chang, "Auditory spectral integration in the perception of diphthongal vowels," in *Journal of Acoustical Society of America*, vol. 128, no.4, pp. 2070–2074, 2010.

[10] S. Y. Yu, Y. D. Chen and J. R. Wu, "Spectral integration and perception of Chinese back vowel γ," in *Science China Information Sciences*, vol. 53, no. 11, pp. 2300–2309, 2010.

[11] G. Peterson and H. Barney, "Control methods used in a study of the vowels," in *Journal of Acoustical Society of America*, vol. 24, no. 2, pp. 175–184, 1952.

[12] "Peterson Barney: Vowel formant frequency database" URL: http://www.cs.cmu.edu/Groups/AI/areas/speech/database/pb/ (Last visited 21st February, 2017)

[13] A. Syrdal and H. Gopal, "A perceptual model of vowel recognition based on the auditory representation of American English vowels," in *Journal of Acoustical Society of America*, vol. 79, no. 4, pp. 1086–1110, 1986.

[14] K. Johnson, "Higher formant normalization results from auditory integration of F2 and F3," in *Perception and Psychophysics*, vol. 46, no. 2, pp. 174–180, 1989.

[15] P. Escudier, J. L. Schwartz and M. Boulogne, "Perception of stationary vowels: internal representation of the formants in the auditory system and two-formant models," in *Franco-Swedish Seminar, Société Francaise d' Acoustique (SFA)*, pp. 143–174 1985.

[16] J. L. Schwartz and P. Escudier, "A strong evidence for the existence of a large-scale integrated spectral representation in vowel perception," in *Speech Communication*, vol. 8, no. 3, pp. 235–259, 1989.

[17] T. V. Ananthapadmanabha, A. G. Ramakrishnan and S. Sharma, "Significance of the levels of spectral valleys with application to front/back distinction of vowel sounds", https://arxiv.org/abs/1506.04828, 2015.