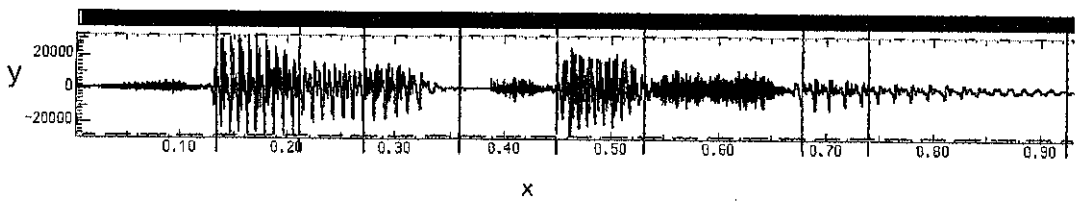
บทที่ 2
หลักการและทฤษฎีที่ใช้

2.1 ความรู้พื้นฐานเกี่ยวกับเสียง

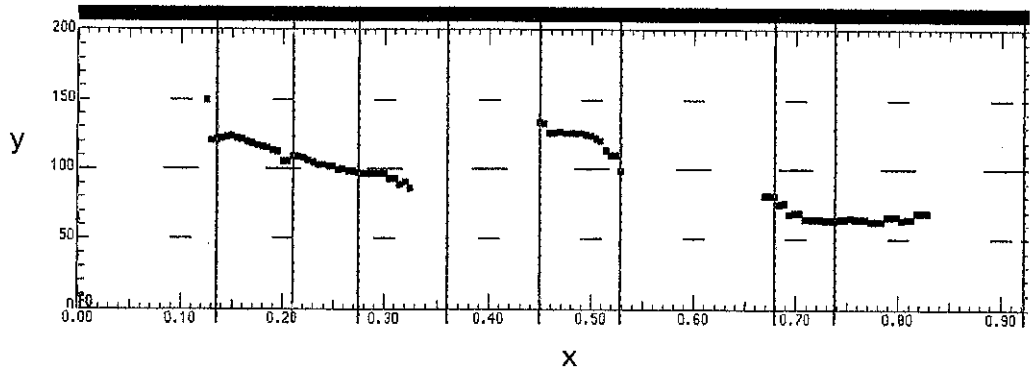
ลักษณะทางกายภาพของสัญญาณเสียงพูด (Speech signal) คือ อนุกรมของการเปลี่ยนแปลงความกดอากาศในตัวกลางระหว่างแหล่งกำเนิดเสียงและผู้ที่รับฟัง ส่วนใหญ่เราจะแทนสัญญาณเสียงด้วย Oscillogram หรือที่เรียกกันว่า Waveform [2] ดังรูปที่ 2.1



รูปที่ 2.1 แสดง waveform

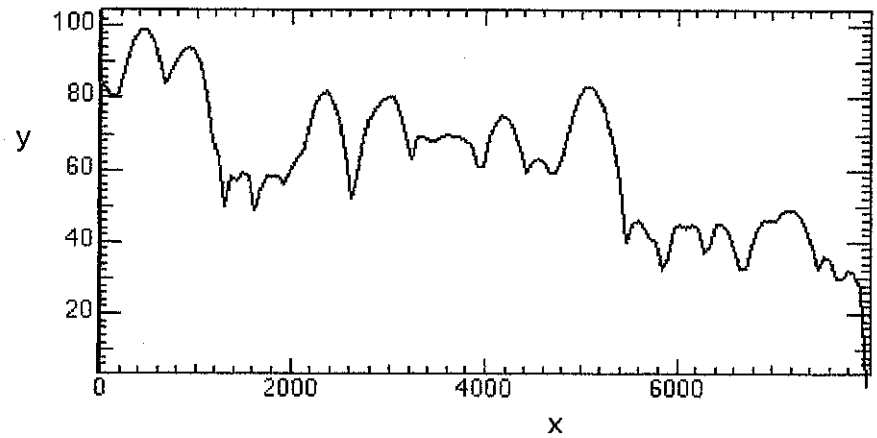
ในแกน x แสดงถึงความกดอากาศที่เปลี่ยนแปลงไปโดยมีการเพิ่มและลดของสัญญาณ สัญญาณเสียงพูดสามารถแทนการวิเคราะห์ระดับเสียง เสียงพูดส่วนใหญ่ประกอบได้สองส่วน คือ เสียงที่เปล่งออก และเสียงที่กรองโดย ฟัน ลิ้น ในปาก การวิเคราะห์ระดับเสียงเป็นสิ่งที่เราพยายามที่จะศึกษาเพื่อให้ได้ทราบถึงความถี่ที่แท้จริงของแหล่งกำเนิดเสียงโดยการวิเคราะห์เสียงพูด ความถี่เสียงที่แท้จริง คือ ขอบเขตของความถี่ในเสียง การวิเคราะห์นั้นไม่ยากต่อการกระทำ สามารถทำได้หลายวิธี เช่น air-flow tube ,electromyography เป็นต้น แต่ปัญหาที่พบในการตัดสินว่าส่วนใช่ต่างๆของสัญญาณเสียงหรือไม่ ยังคงเป็นสิ่งที่ยากต่อการถอดรหัสของสัญญาณเสียง ยังมีการพยายามหาการแก่วงของสัญญาณจากจุดเริ่มต้นของสัญญาณเสียงซึ่งจะถูกกรองจากอวัยวะต่างๆ ในปากของเรา แนวความคิดในการพัฒนา ไม่มีวิธีการใดที่ให้ค่าที่ถูกต้องแน่นอนและมีประสิทธิภาพที่ดีที่สุด ความถี่จึงเป็นพื้นฐานที่สำคัญที่มีความเกี่ยวพันกับความเข้าของผู้ฟังที่จะเข้าใจถึงการเปล่งเสียงพูด

ในภาพที่ 2.2 แสดงถึงความถี่ของสัญญาณเสียงของเสียงผู้ชายและผู้หญิง ซึ่งความถี่ที่ร่วมกันซึ่งเท่ากับ 100 Hz และ 150 Hz โดยช่วงความถี่ของสัญญาณเสียงของผู้ชายจะมีขอบเขตอยู่ระหว่าง 80 – 200 Hz แต่สำหรับผู้หญิงช่วงความถี่ของสัญญาณเสียงจะมีขอบเขตอยู่ระหว่าง 150 -350 Hz



รูปที่ 2.2 แสดงคามถี่ของสัญญาณเสียง

โดยทั่วไปแล้วคาบแต่ละคาบของ Waveform จะเป็นผลบวกของคลื่น sine wave ซึ่งเฉพาะแต่ละ แอมพิจูด ความถี่ และ มุม รูปที่ 2.3 แสดงสเปกตรัมของ waveform ของรูปที่ 2.1 ที่เป็นการกระจายส่วนของ ความถี่และแอมพิจูด ที่เวลาต่าง ๆ โดยที่แกน x เป็นเวลา แกน y เป็นแอมพิจูด ณ ช่วงเวลา 0.15 วินาที ของเสียง

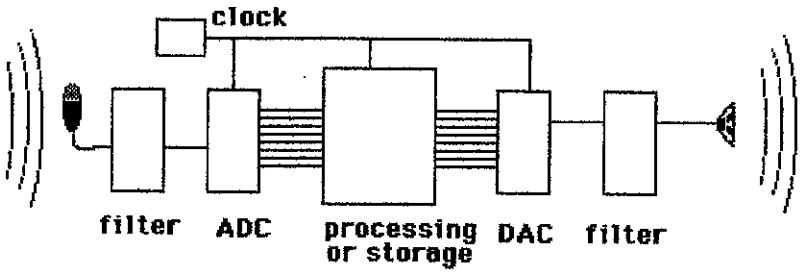


รูปที่ 2.3 สเปกตรัมของ Waveform

2.2 การเก็บข้อมูลเสียง

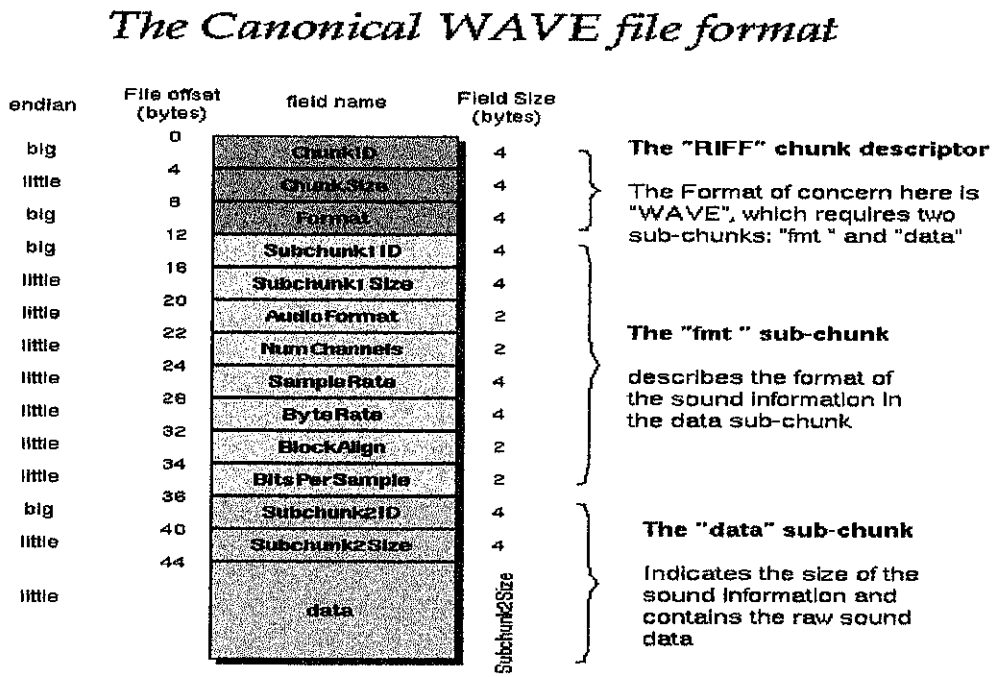
การบันทึกเสียงในปัจจุบันส่วนมากจะอยู่ในรูปแบบของ PCM (Pulse Code Modulation) ซึ่ง PCM ถูกใช้ในการบันทึก CD และ ไฟล์ .WAV ส่วนมากในเครื่องบันทึกแบบ PCM จะแปลงแรงของอากาศให้เป็นแรงดันทางไฟฟ้า จากนั้นตัว A/D จะวัดแรงดันที่ความถี่ในช่วงเวลาหนึ่ง ๆ เช่น การบันทึก CD จะกำหนดให้ค่า Samples อยู่ที่ 44,100 Hz แรงดันนี้จะถูกแปลงเป็นเลขแบบ 16

บิต โดย CD จะเก็บไว้เป็น 2 ช่องสัญญาณคือ ช่องสัญญาณซ้าย และช่องสัญญาณขวา เรียกว่า Stereo ข้อมูลที่ได้จาก PCM จะถูกเก็บในรูปของฟังก์ชันเวลาที่เรียกว่าwave file



รูปที่ 2.4 การบันทึกเสียงแบบ PCM

การเก็บไฟล์แบบ Wave File เป็นการเก็บข้อมูลมัลติมีเดียไฟล์ของ Microsoft โดยที่ไฟล์ RIFF จะเริ่มจาก Header ของไฟล์และตามมาด้วยก้อนของข้อมูล ไฟล์ wave จะประกอบด้วย 2 ส่วนคือ ข้อมูลของ fmt เป็นตัวบอกลักษณะพิเศษของข้อมูล และก้อนของข้อมูลที่บรรจุข้อมูลตามสภาพจริง โดยเรียกแบบนี้ว่า"Canonical form "



รูปที่ 2.5 แสดงการเก็บข้อมูลแบบ Wave file

ตารางที่ 2.1 รูปแบบการเก็บ Wave file โดยเริ่มที่ RIFF Header

Offset	Size	Name	Desscription
0	4	ChunkID	บรรจุตัวอักษร "RIEF" เป็นรหัสแอสกี (0X5249646 big-endian form)
4	4	ChunkSize	36+SubChunk2Size หรือ 4+(8+SubChunk1Size)+((8+Chunk1Size)) ขนาดของข้อมูลที่เป็นตัวเลข ที่ขนาดของไฟล์ที่มี 8 ไบต์สำหรับ 2 ขนาดที่ไม่นำมาในการนับคือ ChunkID และ ChunkSize
8	4	Format	บรรจุตัวอักษร "WAVE" เป็นรหัสแอสกี (0X57415645big-endian form)
12	4	Subchunk1ID	บรรจุตัวอักษร "fmt" เป็นรหัสแอสกี (0X666d7420 big-endian form)
16	4	Subchunk1Size	16 สำหรับ PCM ขนาดที่ได้จาก Subchunk ที่เป็นตัวเลข
20	2	AudioFormat	PCM=1 ค่าของแสดงการบีบอัด
22	2	NumChanels	Mono=1,Stereo=2
24	4	SampleRate	8000,44100,... อัตราความถี่ที่ใช้ในการอัดเสียง
28	4	ByteRate	มีค่าเท่ากับ SampleRate×NumChanels×BitPerSample/8
32	2	BlockAlign	มีค่าเท่ากับ NumChanels× BitPerSample/8
34	2	BitPerSample	8 บิต =8, 16 บิต =16
36	4	Subchunk2ID	บรรจุตัวอักษร "data" เป็นรหัสแอสกี (0x64617461 big-endian form)
40	4	Subchunk2size	มีค่าเท่ากับ NumSample×NumChanels×BitPerSample/8
44	*	Data	ข้อมูล

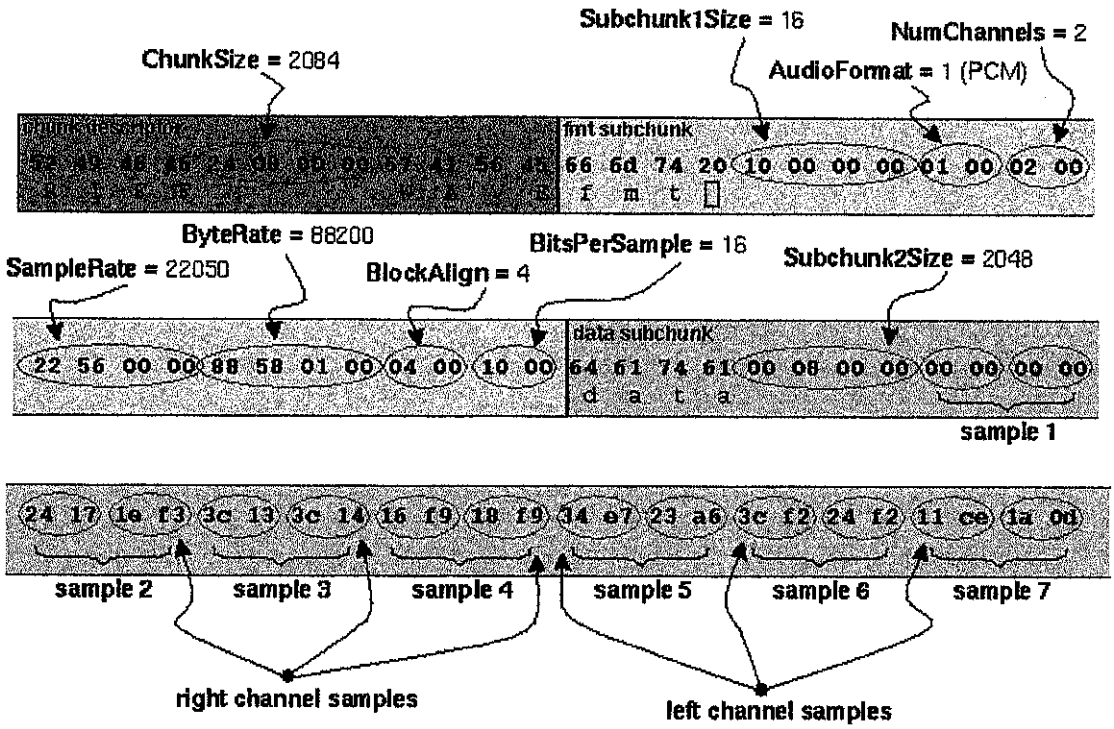
ตัวอย่างมีการเปิด Wave file ที่มีขนาด 72 ไบต์ ที่อยู่ในเลขฐาน 16 ดังนี้

52 49 46 46 24 08 00 00 57 41 56 45 66 6d 74 20 10 00 00 00 01 00 02 00

22 56 00 00 88 58 01 00 04 00 10 00 64 61 74 61 00 08 00 00 00 00 00 00

24 17 1e f3 3c 13 3c 14 16 f9 18 f9 34 e7 23 a6 3c f2 24 f2 11 ce 1a 0d

จากรูปที่ 2.6 จะเป็นการ แสดงการอธิบายข้อมูลแต่ละไบต์ที่เก็บในรูปแบบของ Wave file จากข้อมูลด้านบนที่มีทั้งหมด 72 ไบต์



รูปที่ 2.6 แสดงการอธิบายข้อมูลแต่ละไบต์ที่เก็บในรูปแบบของ Wave file

2.3 ทฤษฎี Cross Correlation Function สำหรับการวิเคราะห์เสียง

จากทฤษฎี Cross Correlation Function [1] ที่นำมาใช้ในการหาความสัมพันธ์ของสัญญาณเสียงของบุคคลแต่ละบุคคลนั้น สมการที่ใช้ในการวิเคราะห์เสียงในโครงงานนี้ เราได้ใช้สมการดังแสดงในสมการที่ (1)

$$\hat{R}_{xy}(L) = \frac{1}{N-L} \sum_{n=1}^{N-L} [x(n+L) * y(n)] \qquad (1)$$

โดยที่

$\hat{R}_{xy}(L)$ คือ ค่าความสัมพันธ์ของสัญญาณเสียง $x(n)$ กับ $y(n)$

N คือ ความยาวทั้งหมดของสัญญาณเสียงในหนึ่งเสียง

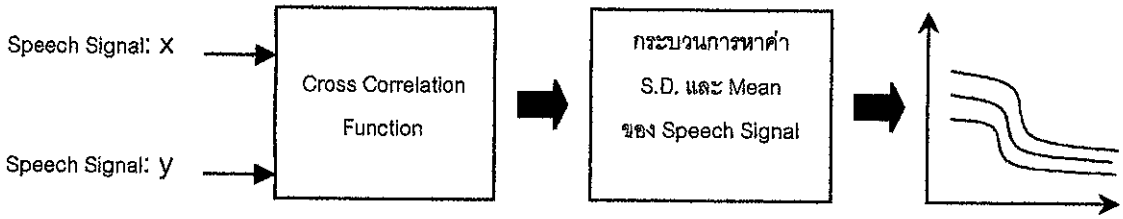
L คือ จำนวนค่าความสัมพันธ์ที่ต้องการหา

$x(n)$ คือ ข้อมูลสัญญาณเสียง x

$y(n)$ คือ ข้อมูลสัญญาณเสียง y

สำหรับในโครงงานนี้ค่า L ที่ใช้มีค่าเท่ากับ 100 ดังนั้นค่าความสัมพันธ์ของสัญญาณเสียงที่เราใช้ทดสอบหรือค่า $\hat{R}_{xy}(L)$ จึงค่าทั้งหมดเป็น 100 ค่า

2.4 Model สำหรับการตัดสินใจเกี่ยวกับเสียง



อธิบายการทำงานของ Model สำหรับการตัดสินใจเกี่ยวกับเสียง จะมีสัญญาณเสียงสองเสียงที่ได้มาการบันทึกเสียงที่อยู่ในรูปแบบของ Wave file ซึ่งสัญญาณเสียง x จะเป็นสัญญาณเสียงที่เราต้องการหาว่าเป็นเสียงของบุคคลใด และสัญญาณเสียง y จะเป็นสัญญาณเสียงที่เราทราบว่าเป็นเสียงของบุคคลใด ซึ่งจะใช้เป็นตัวเปรียบเทียบและหาความสัมพันธ์ของเสียงกับสัญญาณเสียง x โดยผ่านสัญญาณเสียงทั้งสองเข้าในกระบวนการของ Cross Correlation Function จากนั้นจะได้ค่าความสัมพันธ์ของเสียง เมื่อได้ความสัมพันธ์ของเสียงแล้ว จะนำค่าความสัมพันธ์ของเสียง ณ ตำแหน่งต่าง ๆ ของแต่ละตำแหน่งมาหาค่าเบี่ยงเบนมาตรฐาน (Standard deviation) และ ค่าเฉลี่ยเลขคณิต (Mean) ของเสียง ณ ตำแหน่งต่าง ๆ เมื่อผ่านกระบวนการนี้แล้ว เราจะได้ค่าค่าเบี่ยงเบนมาตรฐานและค่าเฉลี่ยเลขคณิต [4] จากนั้นนำค่าที่มาฟล็อตกราฟเพื่อเป็นตัวกำหนดขอบเขตในการตัดสินใจ โดยกราฟจะประกอบด้วยเส้นกราฟ 3 เส้น คือ

1. เส้นค่าขอบเขตบน = ค่าเฉลี่ยเลขคณิต + [FD × ค่าเบี่ยงเบนมาตรฐาน]
2. เส้นค่าขอบเขตล่าง = ค่าเฉลี่ยเลขคณิต - [FD × ค่าเบี่ยงเบนมาตรฐาน]
3. เส้นค่าเฉลี่ย

โดยที่

FD คือ แฟกเตอร์ของการตัดสินใจ

โดยที่ค่า FD สามารถหาได้จากหาค่าสัมพันธ์ของสัญญาณเสียงที่เราทราบว่าใครเป็นเจ้าของเสียง จำนวนหลาย ๆ สัญญาณเสียงแล้วทำการปรับค่า FD ที่ต่ำที่สุดที่สามารถทำให้ ค่าขอบเขตบนและค่าขอบเขตล่าง ครอบคลุมค่าความสัมพันธ์เสียงทั้งหมด จะได้ค่าแฟกเตอร์ของการตัดสินใจถ้าสัญญาณเสียงที่นำมาเปรียบเทียบมีค่าความสัมพันธ์ของสัญญาณเสียงไม่อยู่ในค่าขอบเขตบนและค่าขอบเขตล่างของค่าความสัมพันธ์ของสัญญาณเสียง ก็สามารถบอกได้ว่าเสียงไม่ใช่เสียงของบุคคลคนเดียวกัน แต่ถ้าอยู่ในค่าขอบเขตบนและค่าขอบเขตล่างของค่าความสัมพันธ์ของสัญญาณเสียง ก็สามารถบอกได้ว่าเสียงใช่เสียงของบุคคลคนเดียวกัน

กระบวนการในการหาค่าเบี่ยงเบนมาตรฐานและค่าเฉลี่ยเลขคณิต ของสัญญาณเสียง ณ ตำแหน่งเวลาเดียวกันของสัญญาณเสียง จะทำโดยการอ่านข้อมูลของสัญญาณเสียงจากการบันเสียงที่เป็นแบบ Wave File ข้อมูล ณ ตำแหน่งเวลาต่าง ๆ มาเก็บเป็นข้อมูลแบบอาร์เรย์ (Array Data Type) ดังนั้น ณ ตำแหน่งเวลาเดียวกันของข้อมูลจะถูกแทนด้วย อินเด็กซ์ (Index) ของอาร์เรย์แต่ละอาร์เรย์ การอ้างถึงข้อมูล ณ ตำแหน่งเวลาต่าง ๆ ดังนี้ กำหนดให้สัญญาณเสียงที่ถูกอ่านมาเก็บอาร์เรย์มีทั้งหมด 10 ข้อมูล ดังต่อไปนี้

Index	1	2	3	4	5	6	7	8	9	10
Data	**0.1**	**0.2**	**0.5**	**0.5**	**0.9**	**0.4**	**0.5**	**0.7**	**0.3**	**0.2**

ดังนั้นการอ้างข้อมูล ณ ตำแหน่งเวลาต่าง ก็สามารถอ้างได้จากอินเด็กซ์ของอาร์เรย์ข้อมูลสัญญาณเสียงนั้น จากด้านบน เราสามารถอ้างข้อมูลที่มีค่า 0.9 โดยใช้อินเด็กซ์ 5 ในอ้างถึงข้อมูลดังกล่าว ในการอ้างที่เป็นตัวแปรก็สารมารถทำได้ จากข้อมูลด้านกำหนดให้เท่ากับตัวแปร x(n) ดังนั้นค่า 0.9 ก็จะอ้างได้โดย x(5) ดังนั้นการค่าเฉลี่ยเลขคณิต ของสัญญาณเสียง ณ ตำแหน่งเวลาเดียวกัน สามารถหาได้จากสมการ

$$\overline{X}(k)=\frac{\sum_{i=1}^{N} x_i(k)}{N}$$

โดยที่

$\overline{X}(k)$ คือ ค่า Mean ที่ตำแหน่งที่ k ของ $x_i(k)$

N คือ จำนวนข้อมูลทั้งหมดของ $x_i(k)$

ตัวอย่างในการหาค่าเฉลี่ยเลขคณิต ที่มีข้อมูลจำนวน 3 ข้อมูล คือ x_1 x_2 และ x_3 ในแต่ละข้อมูลมีความยาวข้อมูลเท่ากับ 5 จากสมการด้านบนจะได้ $N=5$ ดังนั้นค่าเฉลี่ยเลขคณิตของข้อมูลจะได้ดังนี้

$$x_1 = [1 \quad 2 \quad 5 \quad 6 \quad 9]$$
$$x_2 = [7 \quad 2 \quad 4 \quad 1 \quad 3]$$
$$x_3 = [5 \quad 4 \quad 6 \quad 3 \quad 4]$$

วิธีการหาค่าเฉลี่ยเลขคณิต ณ ตำแหน่งเวลาเดียวกัน

$$\overline{X}(1) = \frac{x_1(1)+x_2(1)+x_3(1)}{3} = \frac{1+7+5}{3} = \frac{13}{3}$$
$$\overline{X}(2) = \frac{x_1(2)+x_2(2)+x_3(2)}{3} = \frac{2+2+4}{3} = \frac{8}{3}$$
$$\overline{X}(3) = \frac{x_1(3)+x_2(3)+x_3(3)}{3} = \frac{5+4+6}{3} = 5$$
$$\overline{X}(4) = \frac{x_1(4)+x_2(4)+x_3(4)}{3} = \frac{6+1+3}{3} = \frac{10}{3}$$
$$\overline{X}(5) = \frac{x_1(5)+x_2(5)+x_3(5)}{3} = \frac{9+3+4}{3} = \frac{16}{3}$$

ดังนั้น

$$\overline{X} = \left[\frac{13}{3} \quad \frac{8}{3} \quad 5 \quad \frac{10}{3} \quad \frac{16}{3}\right]$$

และในทำนองเดียวกันการหาค่าเบี่ยงเบนมาตรฐานก็สามารถหาได้จากสมการ

$$S.D(i) = \sqrt{\frac{\sum(x_i(i)-\overline{x}_i(i))^2}{n-1}}$$

โดยที่

S.D คือ ค่าเบี่ยงเบนมาตรฐาน

x_i คือ ข้อมูลแต่ละตัว

$\bar{x}_i$ คือ ค่าเฉลี่ยของข้อมูล

คือ จำนวนข้อมูล

จากสมการของค่าเบี่ยงเบนมาตรฐาน และข้อมูลด้านบน ต่อไปจะแสดงวิธีการหาค่าเบี่ยงเบนมาตรฐาน ณ ตำแหน่งเวลาเดียวกัน ดังนี้

$$S.D(1) = \sqrt{\frac{(1-\frac{13}{3})^2+(7-\frac{13}{3})^2+(5-\frac{13}{3})^2}{2}} = 3.0551$$

$$S.D(2) = \sqrt{\frac{(2-\frac{8}{3})^2+(2-\frac{8}{3})^2+(4-\frac{8}{3})^2}{2}} = 1.1547$$

$$S.D(3) = \sqrt{\frac{(5-5)^2+(4-5)^2+(6-5)^2}{2}} = 1$$

$$S.D(4) = \sqrt{\frac{(6-\frac{10}{3})^2+(1-\frac{10}{3})^2+(3-\frac{10}{3})^2}{2}} = 2.5166$$

$$S.D(5) = \sqrt{\frac{(9-\frac{16}{3})^2+(3-\frac{16}{3})^2+(4-\frac{16}{3})^2}{2}} = 3.2146$$

ดังนั้นค่าเบี่ยงเบนของข้อมูลดังกล่าวจะได้เท่ากับ

$$S.D=[3.0551 \quad 1.1547 \quad 1 \quad 2.5166 \quad 3.2146]$$