Japanese Speech-to-Text system

Akira Morimoto *

Keywords:Japanese Speech-to-Text system, Wavelet transform, Growing cell structure

Abstract

In this paper, the Japanese speech-to-text system is built. By wavelet transformation, Japanese speech changes into more distinguishing time-frequency information than the original signal. The patterns of time-frequency which often appear in Japanese speech are extracted by using the growing cell structure. The growing cell structure is trained by the ordinary Japanese speech for several hours. The basic syllable file makes the growing cell structure to be reduced, and it is possible to transform Japanese speech to text. Finally, an experiment result is described.

1 Introduction

In this research, Japanese speech-to-text conversion system is constructed. The present stage is performing the training and the recognition of a growing cell structure network only by the author's speech.

The section 2 describes the features of Japanese speech. In the section 3, continuous wavelet transformation is used in order to take out time-frequency information from a speech signal. The following section 4 describes the growing cell structure which extracts the features from the time-frequency information obtained by wavelet transformation. In the section 5, the growing cell structure is made to reduce by using a basic syllable file, then Japanese speech recognition is performed. The final section 6 reports an experiment result, and the problems and the future research subject are mentioned.

2 The features of Japanese speech

Japanese has a phonetic syllabary which is called "kana". The "kana" alphabet and its reading correspond one-to-one. The Japanese syllabary is raised to table 1. The number of Japanese syllabaries is about 100. The "kana" alphabets which cannot distinguish by pronunciation was bundled with 〔 〕.

When Japanese phones are divided into vowels and consonants, Japanese has only 5 kinds of vowel sounds (あ(a)/い(i)/う(u)/え(e)/お(o)). Moreover, the reading is hardly influenced by the order of syllabaries. The reading of syllabary changes a little according to the last syllabary and next syllabary. If it is in English or French case, the pronunciation will be influenced of about seven alphabets order.

Since the intonation of a sentence is also few, and since one phone corresponds to one "kana" alphabet, time-frequency analysis is very effective in the Japanese speech-to-text conversion system. In this paper, the Japanese speech is a monophonic recording , 16 bits per one sample, 8000 sampling per second sound.

*Osaka kyoiku University

あ (a)	い (i)	う (u)	え (e)	お (o)
か (ka)	き (ki)	く (ku)	け (ke)	こ (ko)
さ (sa)	し (si)	す (su)	せ (se)	そ (so)
た (ta)	ち (ti)	つ (tu)	て (te)	と (to)
な (na)	に (ni)	ぬ (nu)	ね (ne)	の (no)
は (ha)	ひ (hi)	ふ (fu)	へ (he)	ほ (ho)
ま (ma)	み (mi)	む (mu)	め (me)	も (mo)
や (ya)		ゆ (yu)		よ (yo)
ら (ra)	り (ri)	る (ru)	れ (re)	ろ (ro)
わ (wa)	［うぃ (wi)］		［うぇ (we)］	［を (wo)］
が (ga)	ぎ (gi)	ぐ (gu)	げ (ge)	ご (go)
ざ (za)	じ (zi)	ず (zu)	ぜ (ze)	ぞ (zo)
だ (da)	［ぢ (di)］	［づ (du)］	で (de)	ど (do)
ば (ba)	び (bi)	ぶ (bu)	べ (be)	ぼ (bo)
ぱ (pa)	ぴ (pi)	ぷ (pu)	ぺ (pe)	ぽ (po)
きゃ (kya)	［きぃ (kyi)］	きゅ (kyu)	［きぇ (kye)］	きょ (kyo)
ぎゃ (gya)	［ぎぃ (gyi)］	ぎゅ (gyu)	［ぎぇ (gye)］	ぎょ (gyo)
しゃ (sya)	［しぃ (syi)］	しゅ (syu)	［しぇ (sye)］	しょ (syo)
じゃ (zya)	［じぃ (zyi)］	じゅ (zyu)	［じぇ (zye)］	じょ (zyo)
ちゃ (tya)	［ちぃ (tyi)］	ちゅ (tyu)	［ちぇ (tye)］	ちょ (tyo)
［ぢゃ (dya)］	［ぢぃ (dyi)］	［ぢゅ (dyu)］	［ぢぇ (dye)］	［ぢょ (dyo)］
にゃ (nya)	［にぃ (nyi)］	にゅ (nyu)	［にぇ (nye)］	にょ (nyo)
ひゃ (hya)	［ひぃ (hyi)］	ひゅ (hyu)	［ひぇ (hye)］	ひょ (hyo)
びゃ (bya)	［びぃ (byi)］	びゅ (byu)	［びぇ (bye)］	びょ (byo)
ぴゃ (pya)	［ぴぃ (pyi)］	ぴゅ (pyu)	［ぴぇ (pye)］	ぴょ (pyo)
みゃ(mya)	［みぃ (myi)］	みゅ (myu)	［みぇ (mye)］	みょ (myo)
りゃ (rya)	［りぃ (ryi)］	りゅ (ryu)	［りぇ (rye)］	りょ (ryo)
ん(nn)	っ (tt)			

Table 1: Japanese phonetic syllabary

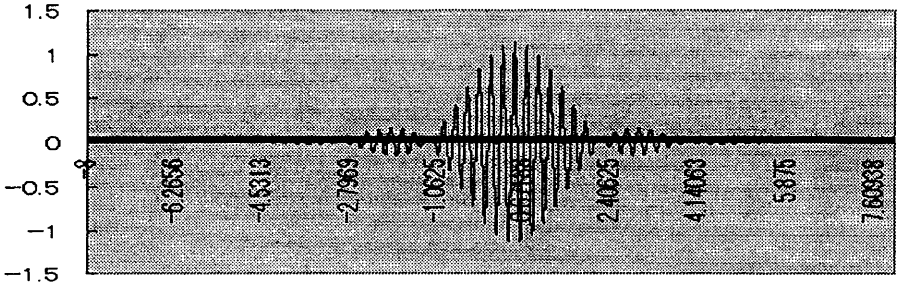

Figure 1: Meyer wavelet $N = 8$

scale j	main frequency	time step	time step of the orthonormal basis
33	184 Hz	50/8000 sec = 0.00625	0.021 sec
34	206 Hz	50/8000 sec = 0.00625	0.018 sec
...	...	...	...
42	530 Hz	50/8000 sec = 0.00625	0.0071 sec

Table 2: The discretization of parameters and main frequency

3 Continuous wavelet transformation

In this paper, a speech signal is not treated as it is, but we deal with the time-frequency information obtained from the speech signal using continuous wavelet transformation. Here, we use the Meyer wavelet with rational dilation factor ([1],[2]).

For a natural number $N \in \mathbf{N}$, this wavelet $\psi(x)$ has the dilation factor $\alpha = \frac{N+1}{N}$, such that the set of functions

$$\psi^{(j,k)} = \alpha^{j/2}\psi\left(\alpha^j x - k\right), \quad j,k \in \mathbf{Z}, \tag{1}$$

constitutes an orthonormal basis of $L^2(\mathbf{R})$. We choose $N = 8, \alpha = 1.125, \alpha^6 = 2.027 \sim 2$ (see figure 1). The main frequency of this wavelet $\psi(x)$ is about $\frac{2N^2}{1+2N}\pi = (128/17)\pi \sim 7.53\pi = 7.53/2$Hz.

The discretization of the parameters is

$$\alpha^{j/2}\psi\left(\alpha^j(x - kB)\right), \quad j,k \in \mathbf{Z}, \tag{2}$$

and we choose $\alpha = 9/8, B = 50/8000 = 0.00625$. Furthermore, the scale parameter $j = 33, 34, \cdots, 42$ are only used (table 2). The main frequencies of corresponding to these scales are from 184 Hz to 530 Hz. As processing of a male voice signal, although the larger frequency band (from 100 Hz to 1000 Hz) should have been investigated, in order to shorten calculation time, the frequency band was narrowed down.

By continuous wavelet transformation, a speech signal of 8000 samplings per second is changed into the time-frequency information which consists of 240 10-dimensional vectors per second. As the example, figure 2 shows the voice signal which the author pronounced Japanese vowel "あ(a)" ,and the wavelet transform corresponding to this vowel is raised to figure 3. This continuous wavelet transform

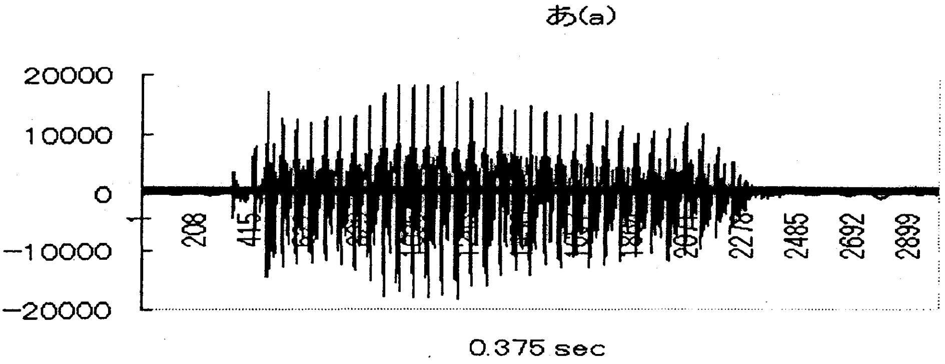

Figure 2: A voice signal of Japanese vowel "あ(a)"

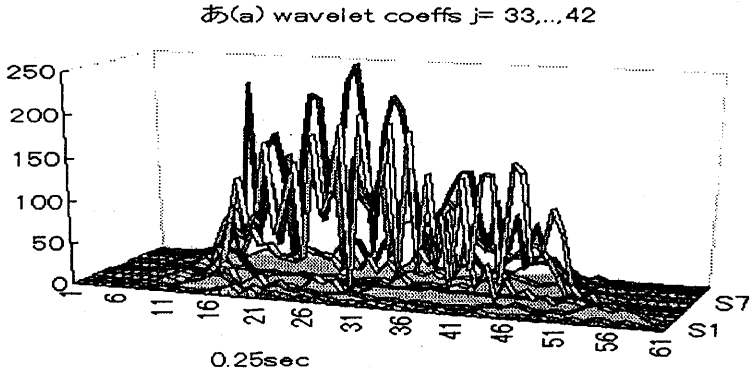

Figure 3: The absolute value of wavelet coefficients corresponding to "あ(a)". Frequency is so high that it becomes this side.

is calculated by numerical integration, assumed that this wavelet's support is $[-3.5, 4.5]$. Moreover, the only absolute value of wavelet coefficients is used in this research.

4 Growing cell structure

In the foregoing section, using continuous wavelet transformation, the speech signal was changed into the time-frequency information. This time-frequency information is a time series of 240 10-dimensional vectors per second. Since the author usually pronounce 1 Japanese phonetic syllabary from 0.2 seconds to 0.5 seconds, 1 syllabary has from 50 to 100 10-dimensional vectors.

What phonetic syllabary is in signal is presumed by along what orbit these 10-dimension vectors pass. For that , it is necessary to pursue or approximate these 10-dimention orbits. Here, we use the growing cell structure proposed by B. Fritzke([3]).

The growing cell structure is the graph which consists of vertexes and edges in 10-dimensional space.

The structure is changed according to the inputs. Each vertex $\{v_i\}$ has 2 variables. One is the position of the vertex in 10-dimensional (v_i.position) space, and another is the accumulated error (v_i.error). Each edge $\{e_j\}$ has 4 variables. First 2 variables are the numbers of two vertexes where the edge connects (e_j.vertex_no1,e_j.vertex_no2), 3rd variable is age of the edge (e_j.age), and last variable is the accumulated error of the edge (e_j.error).

First, we take two or more vertexes which have different positions each other. Next, the graph is changed as follows, according to the input w .

1. Chose the point v_k nearest the input w and the point v_l near the next.

2. The distance to the input w is added to the accumulated error of the nearest point (v_k.error).

3. The nearest point is moved in the input direction at a rate of *alpha*, and the point near the next is also moved in the input direction at a rate of *beta* ($\alpha > \beta$).

4. IF there is the edge which connects v_k and v_l, then the age of this edge sets 0, and the distance to the input w is added to the accumulated error of this edge. Else we make a new edge which connects v_k and v_l, set the age of this edge to 0, and let the accumulated error of this edge be the distance to the input.

5. 1 is added to the age of every edge.

For a fixed period of inputs, an addition of a new vertex, deletion of edges, and deletion of points are performed. Therefore the graph is updated.

When the accumulated error of vertex v_k which has the maximum accumulated error exceeds the threshold (error_threshold), an addition of new vertex is performed in the following ways.

1. Select the edge e_j which connects with v_k and which has the maximum edge error. The vertex v_l is another connected vertex of this edge e_j.

2. A new vertex v_m is made at the midpoint of the edge e_j, and the error v_m is equal to the half of the error of this edge. Moreover, the error of v_k and v_l is reduced by the half of the error of this edge.

3. A new edge which connects v_k and v_m, and another new edge which connects v_l and v_m are produced. Let their edge error be the half of the error of edge e_j. Delete the edge e_j.

4. The accumulated error of all vertexes and edges is $gamma(< 1)$ times.

When the age of some edge exceeds the threshold (edge_threshold), this edge is deleted. Also, the vertex which is connected with no edges is deleted.

In this paper, the parameters of the growing cell structure are defined as follows.

$$\alpha = 0.01, \quad \beta = 0.0001, \quad \gamma = 0.5,$$

$$\text{error\_threshold} = 5000.0, \quad \text{edge\_threshold} = 96000.$$

An addition and deletion of a vertex, and deletion of a edge were worked to every 240 inputs (i.e. per 1 second). Moreover, the edge not used during 400 seconds was deleted.

In this parameter setting, after the growing cell structure was trained 3 times for the 3 hours Japanese speech signal, the graph consisted of 5512 vertexes and 71452 edges.

5 Reducing the growing cell structure and pattern recognition

In the foregoing section, when the growing cell structure was trained for the Japanese speech signal, the graph became very large, and so real-time processing became hard. Then the graph will be narrowed down, and the pattern recognition of speech signal will be performed.

In order to learn the correspondence between continuous wavelet transforms of the reading of Japanese phonetic syllabary and Japanese characters, the text file which corresponds with the voice signal is prepared. This voice signal consists of single sounds raised to table 1 and the set of 3 successive single sounds.

The parameters of the growing cell structure is defined as no new vertexes is inserted and as α, β is small enough. Then the growing cell structure is trained for this voice signal, until the number of vertexes and edges becomes fewer about 10 percent.

Next, when the voice signal pass the vertex and edge of the growing cell structure, the character corresponding to the voice signal is multiply assigned to the vertex and edge. Moreover, the silent state where it does not correspond to a voice is registered as a pause of the syllable. Especially the vertexes and edges characteristic of consonants are marked importantly.

In order to make growing cell structure smaller, the vertexes that the distance for 2 vertexes is near are unified at the new vertex. Moreover, the edges are integrated simultaneously. However, in the case of a vertex characteristic of consonants, this operation is not performed. By operation of reducing this graph, the number of vertexes and edges is reduced about to 1/3, and the number of vertexes is made into about 2000 pieces.

The character recognition process works as follows. While scanning input w, until it comes to the vertex corresponding with a character, no speech recognition is carried out. If four edges showing the same character continue, then recognition process starts. Whether the vertex expresses the vowel sound or the consonant sound is distinguished. In the case of consonant, since the consonant is connected with a vowel, when the vowel sound has come, we detect what the syllable is. In a vowel case, a character is decided as the vowel sound as it is. We measure time until the new consonant comes or until another vowel comes or until the silent state comes. If this time interval is less than 0.1 seconds, the detected character is canceled as if it was the error. If this time interval is from 0.2 seconds to 0.5 seconds, the detected character is written in the output text file. If this time interval is very long, it judges with what the same vowel sound follows, and the vowel sound is added suitably.

6 An experiment result and problems

With an author's speech, this speech-to-text system was tried . The specs of the used computer are that CPU is Pentium 120 MB and loading memory is 96 MB. It is 0.5 seconds necessary to convert 1 seconds speech signal to continuous wavelet transform.

A great portion of time required to train the growing cell structure is calculation of the distance between the input and the vertex. When the growing cell structure has 5000 vertexes, it is 5 seconds necessary to train for 1 seconds speech signal. Therefore, it is required for 45 hours to learn 3 times the speech signal of 3 hours.

The process which makes the growing cell structure reduce is required about 6 hours. The work so far doesn't need to be performed on real-time.

Now, we will investigate the required time of speech-to-text process which will be desired to perform on real-time. For 1 seconds speech signal, the wavelet transform needs about 0.5 seconds. Since the growing cell structure has about 2000 vertexes, this graph needs about 2 seconds. The last recognition work is required about 0.5 seconds. After all, it is required for about 3 seconds to recognize 1 second speech signal by the present system.

About recognition accuracy, this system is extremely sensitive to the height of volume, the speech speed and the height of voice. Therefore, if possible, it is necessary to maintain the volume and the pronunciation speed of the same about as the time of learning.

When the speech divided for every syllable is inputted, the rate of a correct answer was 95 % . In this case, the errors are mostly produced by the noise which is blowing the breath on the microphone , or the height of volume.

Next, in the case of speaking ordinarily, the rate of recognition is about 80 %. The most errors occur in a mistake of the pause of characters.

In order to reduce the sensitive nature to volume, when the ℓ^2 norm of wavelet coefficients exceeds a certain threshold below which it is judged that this part of speech is silent, the wavelet coefficients are divided by there ℓ^2 norm. Then, although this system become strong to change of volume, it become difficult to discern consonants which are called fortises (i.e. p,t,k,etc.). For example, "あか(aka):red" or "かた(kata):shoulder" are hard to recognize. The rate of recognition in this case is about 75 % .

A future subject is recognition with another person's speech. First, since from the point of calculation speed, the frequency band was set by the author's speech and narrowed down as much as possible, it needs to extend.

Second, since the work which makes 3 hour speech signal to train the growing cell structure is very tedious, and since calculation time of this training is very long, there is an author's speech in this portion. Then the correspondence file of sounds and characters is created by another man, and speech-to-text test will be performed. In this case, work which shifts a frequency band will be also needed.

References

[1] Y. Meyer,*Principe d'incertitude, bases hilbertiennes et algèbres d'opérateurs,* Séminaire Bourbaki, no. 662, 1985.

[2] P. Auscher, *Wavelet bases for $L^2(\mathrm{R})$, with rational dilation factor,* in Wavelet and there Application , ed. Ruskai et al., Jones and Bartlett, pp. 439-452, 1992.

[3] B. Fritzke,*Growing cell structures – a self organizing network for unsupervised and supervised training,* Neural Networks, no. 9, pp. 1441-1460, 1995.