First European Conference on
Speech Communication and Technology
EUROSPEECH '89
Paris, France, September 27-29, 1989

A 100bit/s Speech Coding using a Speech Recognition Technique

Yoshimitsu Hirata and Seiich Nakagawa

Department of Information and Computer Sciences
Toyohashi University of Technology
Toyohashi, Tenpakucho, 440, Japan

ABSTRUCT

In this paper, we describe a phonetic vocoder based on syllable-units which represents speech waves by extremely low rate (100 bits/s) using a speech recognition tequnique. We take syllables into consideration as the unit of recognition / synthesis. Speech waves are transformed into a sequence of frames, each of which consists of LPC cepstrum, PARCOR coefficients, pitch and power. After the O(n)DP matching with reference patterns, the input speech is transformed into a sequence of Japanese syllables. The information of recognized syllable contains the category of syllables, duration, power and pitch, and is represented by 16 bits. Using this vocoder, speech can be represented by only 100 bits/sec.

1. INTRODUCTION

The extremely low bit rate coding is effective in the reduction of a large quantity of speech data (e.g., computer mail) or mobile communication. For a very low bit rate coding of speech, segment quantization methods based on vector / matrix quantization have been contrived and made good intelligibility at 150～200 bits/s [1][2]. But these methods are regarded as a kind of pattern matching vocoders and do not use language knowledges. That's why such a very low bit coding is in possible for any language. In order to make furthermore low bit coding possible, we must take linguistic knowledges into coding.

There are some studies on such a vocoder, e.g. a phonetic vocoder which finds a phoneme string to minimize the distance between the input speech and diphone templates and which can represent speech at 100 bits/s[3]. Because there is a limitation to the next phoneme by a diphone network, that is, context-dependency, there is not obtained enough intelligibility. It was reported that the intelligibility of this vocoder was improved by the allowance to follow any templates, that was a segment vocoder using diphones as segments (200 bits/s)[4]. Another type of phonetic vocoders was proposed recently[5][6], and they used recognizers based on HMM's to code speech. Picone and Doddington transformed speech into a sequence of phonemes (120 bits/s for spectral information) and durations (50 bits/s for state transitions). This vocoder's quality was comparable with that of a VQ-based vocoder of 300 bits/s. Soong proposed a speech recognition / synthesis method based on 2084 different left and right context-dependent tri-phone model.

In this paper, we discribe a vocoder which can represent speech at 100 bits/s using a speech recognition technique. The synthetic unit of the vocoder is a syllable. Our study is based on the assumption that 1)human beings has high ability of linguistic understanding even if speech recognition by machine is not incomplete or in noisy environments, 2)there is no direct relationship between phoneme recognition accuracy by machine and intelligibility by a speech synthesizer. First, we discribe how speech is coded into an optimal syllable sequence. Next, we report on some experimental results to evaluate this vocoder.

2. CODING INTO AN OPTIMAL SYLLABLE SEQUENCE

2.1 System organization

Considering that the source of speech in

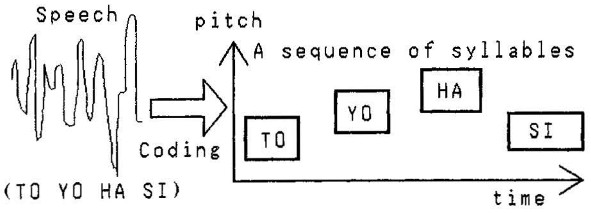

Fig.1 Coding into an Optimal Syllable Sequence

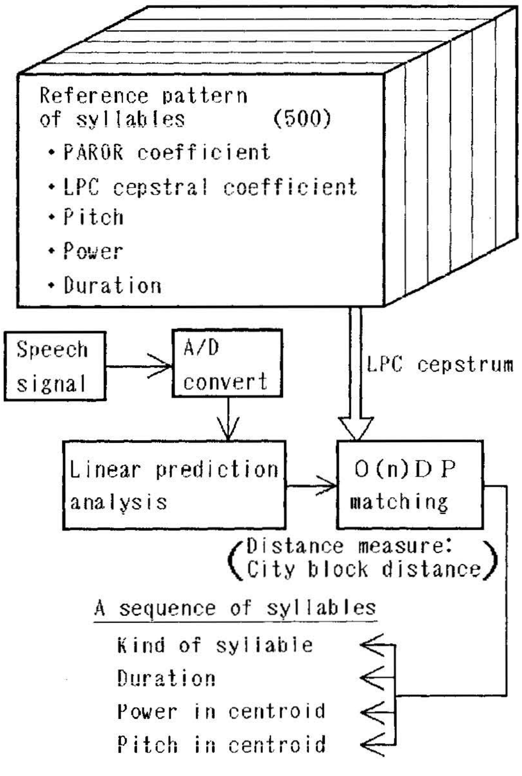

Fig.2 Block Diagram of the Coder

brain is a sequence of discrete symbols, the ultimate coding is obtained by coding into the symbol. There are some units of languege such as phoneme, demi-syllable, syllable, word, etc. We consider syllables as a minimum unit of the language which can be dealt with easy in co-ariculation. The way we use in coding speech is shown in Fig.1. Fig.2 shows a block diagram of the coder.

Input speech waves are sampled at 10kHz in 12 bits with an A/D converter and transformed into a sequence of frames by the 14th oder linear prediction analysis. Each frame has LPC cepstral coefficients, PARCOR coefficients, pitch and power. After the O(n)DP[7] (or One Stage DP[8]) matching with reference patterns, the input speech is transformed into a sequence of syllables. A recognized syllable contains the category of syllables, duration, power and pitch. Each item is quantized by the levels of 500, 3, 5 and 7, respectively, that is, the spectral information is quantized by 9 bits/syllable and prosodic information 7 bits/syllable. If the speech rate is 6 syllables/sec, the amount of coded speech is 96 bits/sec.

2.2 Reference pattern

It is known that plural patterns are need in each category as reference, because the acoustic property varies in context, in other words, it has allophones. We made reference patterns of about 500 syllables (in Japanese, there are about 100 syllables) from 416 words utterd in isolation. Each syllable corresponds to one CV syllable extracted from all VCV syllables which are included in the 416 words, where C and V denote a consonant and a vowel, respectively. These extraction was performed by hand labeling.

2.3 O(n) DP matching method

For continuous speech recognition, we have used the O(n) DP matching method[7]. The O(n) DP matching algorithm is computationaly more efficient than the Two Level DP matching[9] and make an optimal syllable sequence according to a distance measure.

We use LPC cepstral coefficients as a matching parameter for it is better than PARCOR coefficients from recognation accuracy view. And we use the city block distance (Chebyshev norm) as a spectral distortion measure.

2.4 Sampling and coding of prosody

As we discribed above, after the recognition with the O(n) DP matching, input speech is transformed into a sequence of syllables which have category of syllables and prosodic information. In this section we discribe how to sample and code the information of prosody which consists of pitch, power and duration.

Pitch is represented as one in the centroid, and the value divided by the pitch in the preceding syllable is quantized by 7 levels (100/130, 100/120, 100/110, 100/100, 100/90, 100/80, 100/70). A pitch of centroid is defined as an average of a section from 1/2 to 5/6 of syllable obtained by the recognizer. In this way we need some representation of the first syllable pitch, so it is represented as the ratio to a constant value.

Power is quantized as the ratio of the power in the centroid to the coresponding one of the reference pattern by 5 levels (0.1, 0.5, 1.0, 5.0, 10.0). It is another way to represent the power, that is, the absolute power value is quantized. If we use the latter, we need many of levels to quantize the power, so extremely low bit coding doesn't come true. Assuming the ratio of power to one of the coresponding reference is in small range, it is better to use the former. Because there is a high correlation between power and pitch, it is enough to quantize power at 3 levels. However we have not yet used the correlation.

Duration is represented as the ratio the length of the syllable obtained by matching to that of the coresponding reference pattern, and the ratio is quantized by 3 levels (5/6, 8/6, 11/6).

After all each syllable can be represented by 16 bits, and the speech which duration is one second can be represented by only 100 bits if speech is spoken at a speed of 6 syllables/s.

3. DECODING METHOD

Fig.3 shows the block diagram of decoder. For a sequence of quantized codes which are obtained by a recognizer, speech is decoded by concatenating the corresponding reference patterns. We use a CV-concatenation method essentialy and linear-interpolate PARCOR coefficients between syllables. Because each CV sylable has several reference patterns, the reference patterns corresponding to recognition result are adopted. Assuming that input speech is "TO YO HA SI", for example, a desired reconition result may be "TO $_{A}$YO $_{U}$TA $_{A}$SI" ("$_{A}$YO" means a CV syllable "YO" which is a part of VCV triphone "AYO"). Considering the result is optimal from the view to minimize spectral distortion, it is unnecessary to recognize as "TO $_{O}$YO $_{O}$HA $_{A}$SI". And it is unnecessary to correct a syllable sequence because the decoded speech has high ability to be understood as the origin by a human perceptibility if the spectral distortion is remarkably small.

Duration in decoding is modified by eliminating or repeating a part of the vowel (a section from 1/2 to 5/6 of a syllable) of the reference pattern, for it may be hear as another phone if a consonant part is warped.

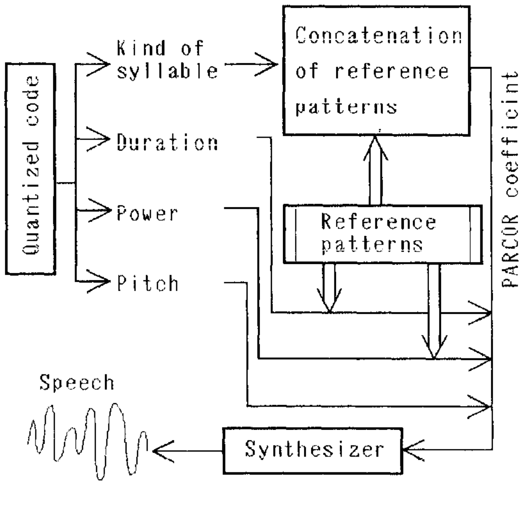

Fig.3 Decoding Method

4. EXPERIMENTS

In a preliminary experiment, we tested the intelligibility of this vocoder for limited 100 spoken words. The subjects identified the presented word at correct rate of about 96% when it was tought in advance that the word was one of the list of 100 words. In this section we describe several expriments in order to test the feasibility for sentences of the vocoder.

First, the intelligibility of synthetic voice was compared with the case of a VQ-based vocoder and without segmentation error in the syllable recognition. Secondly, the pitch information of speech was added in decoding. Finally, we tried to convert voices to a standard speaker's one who was different from a person in coding.

4.1 In the case of perfect segmentation

In order to assess an upper-bound ability of this vocoder, we checked the intelligibility of the vocoder in the no case of segmentation error. The O(n) DP algorithm was modefied not to make any syllable boundary except them given by manual segmentation. We used some sentences spoken with pauses between phrases at a speed of 6～7 mora (12～14 phonemes) per second. In this experiment, we used about 85 bits/s coding rate except for the pitch information.

Fifteen sentences spoken by a male speaker HN who uttered isolated words for reference patterns were recognized in two way. One was the O(n) DP matching and the other was the matching with given boundaries. The recognition result is shown in Table 1.

Table.1 Syllable Recognition Results

method	O(n)DP	boundary known
syllable recognition rate	53%	69%
insert error rate	25%	0%
deletion error rate	2%	0%
segmentation rate	73%	100%
coding distortion	1.46	1.65

The coding distortion means an average distance normalized by the input frame length on matching according to city block distance. The segmentation rate is defined as follows:

$$\text{Segmentation rate} = \frac{\text{INP} - (\text{INS} + \text{DEL})}{\text{INP}} \quad (1)$$

INP: Number of input syllables
INS: Number of inserted syllables
DEL: Number of deleted syllables

By giving boundaries, the syllable recognition rate was improved from 53% to 69%. The reason why the coding distortion of the O(n)DP matching is smaller than one of boundary known case is that the O(n) DP matching algorithm concatenates the best reference patterns and marks boundaries in input speech to minimize the distance between input speech and any string of reference patterns.

Table 2 shows which level of pattern matching vocoder based on vector quantization correspond to or equivalent to the distortion of the O(n) DP matching, where it is obtained from three sentences "Hana yori dango" ("Cake is prefered to flower" in English), "Migini sanjudo maware" ("Rotate to right by 30 degrees") and "Itokowa sizukana ongakuga totemo sukidesita" ("My first cousin liked soft music").

Table.2 Coding Distortion
(Distance measure:
city block distance of PARCOR coefficients)

codebook size	distortion	spectral information
16 (VQ)	1.52	400 bits
32 (VQ)	1.39	500 bits
64 (VQ)	1.29	600 bits
128 (VQ)	1.20	700 bits
256 (VQ)	1.14	800 bits
O(n)DP	1.54	54 bits

We took place a hearing test in the way as follows: The synthesized voices for fifteen sentences were presented by eight subjects. Each of the sentences is simple but the task is not restricted. Five sentences among the fifteen sentences were synthesized by the syllable vocoder with unknown boundaries(pitch is constant), the same vocoder with known boundaries and a pattern matching vocoder based on vector quantization (codebook size of 64, 600 bits/s), respectively. The subjects who don't know the contents of sentences listened twice per one synthetic sentence, then they dictated sentences at their listening. As a result, the phrase inteligibility was about 60% for every vocoders.

The coding distortion of the O(n) DP matching corresponds to one of the pattern matching based on vector quantization in 16 codebook size. But the inteligibility of the O(n) DP matching vocoder was same as a pattern matching vocoder quantized at 64 levels. In the no-error segmentation, the intelligibility was same as unknown boundary case in spite of out expection that the inteligibillity would be improved by rising the recognition result. The reason is considered as follows: even if the recognition accuracy is not good there may be a littel influence on human perception because a sequence of spectral sequence obtained from the O(n) DP matching preserves the linguistic information, in particular, the dynamic information. This advantage is the same as that of a segment vocoder.

4.2 Addition of pitch information

It is important to use prosodic information to improve the syllable recognition rate or to bring on the naturality. We compared the intelligibility of the vocder by a constant pitch with one of the vocoder by the pseudo-real contour pitch. The pitch period was given for three points per one recognized syllable. Each of them was quantized at 7 levels in the same way described before. In decoding, we linear-interpolated pitches at the point between them.

We prepared 28 sentences and reference patterns spoken by each of male speaker YH and female speaker TM, respectively. The inteligibility of O(n) DP matching vocoder was tested for both cases of the constant pitch and variable pitch. To compare them the same condition was hold for a pattern matching vocoder based on the vector quantization method (code book size is 64 and the coding rate is 600 bits/s.). For sentences spoken by speakers

TM and HN, the number of subjects was eight, for speaker YH, five. Table 3 shows exprimental results. We should notice that the stimulated sentences for each item in experiments were different from one another. Therefore, the intelligibility depends more or less on the contents of sentences. For both male speaker YH and female speaker TM, the inteligibility of the O(n) DP matching vocoder was improved by adding pitch information. But there were no change for speaker HN. Besides the syllable recognition result was about 51% recognition rate, 87% segmentation rate for 28 sentences spoken both speakers YH and TM.

Table.3 Phrase Intelligibility

type	pitch	speaker		
		YH	TM	HN
O(n)DP	constant	55%	53%	59%
	variable	67%	56%	60%
VQ	constant	80%	74%	60%
	variable	99%	93%	—

4.3 Voice conversion to a standard speaker's voice

To put such an extremely low bit rate coding into practice, the decoding speech using a speaker's voice who is same as one in coding is not a good way because of neccesary of transmitting reference patterns. This problem can be avoided by preparing one set of a standard speaker's reference patterns beforehand to concatenate them according to a sequence of syllables which have kinds of syllables and prosodic information.

Fig.4 shows how to decode as a standard speaker's voice. Using recognition results, the reference patterns of the speaker in coding are replaced with those of the standard speaker. Input speech is coded by the method stated previously. On side of receiver, the standard speaker's PARCOR coefficients are concatenated coresponding to kinds of syllables. Pitch is multiplied by a ratio of the average standard speaker's pitch to the average input speaker's pitch.

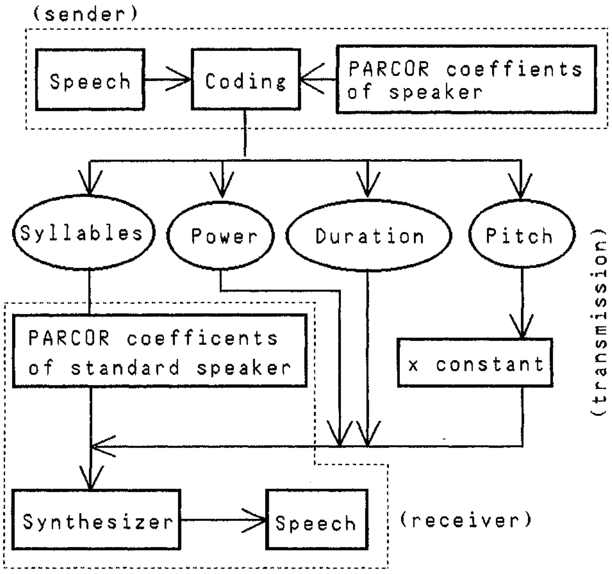

Fig.4 Synthesis as a Standar Speaker's Voice

We tested the intelligibility in the case of converting speaker HN's voice into speaker YH's voice. Five subjects listened to synthetic speech (pitch is constant) twice and dictated what was said. The phrase intelligibility was 46%. Since the original speaker HN's one was about 60% as shown in Table 3 with no pitch information, there was degradation by converting into another speaker's voice. A reason for this result is considered as follows: there is no problem in listening because the matching is done in oder to minimaize distances between input and reference patterns, but the converted spectral sequence may not be optimal in replacing into other speaker's reference patterns.

5. CONCLUSION

We described how to realize a 100 bit/s coding of speech. The purpose of this vocoder is a study of representing speech as a text file. Finally the phrase intelligibility of this vocoder was a little less than 70% in the best case. In order to imporove the intelligibility, following points are considered as effective: a study in increasing reference patterns into around 1000～2000 (10～11 bits to represent a kind of syllables and 110～120 bits/s coding), introduction HMM's to imporove recognition results or to extract automatically reference patterns[10], and representing speaker characteristics in low bit for unspecified speakers.

REFERENCES

[1] S.Roucos, R.M.Schwartz and J.Makhoul: "A Segment Vocoder at 150 B/S", Proc. ICASSP, pp.61-64 (1983).

[2] Y.Shiraki and M.Honda:"LPC Speech Coding based on Variable-Length Segment Quantization", IEEE Trans. Acoust. Speech & Signal process., ASSP-36, 9, pp.1437-1444 (1988).

[3] R.Schwartz, J.Klovstad, J.Makhoul and J.Sorensen:"A Preliminary Design of a Phonetic Vocoder based on a Diphone Model", Proc. ICASSP, pp.32-35 (1980).

[4] S.Roucos, R.Schwartz and J.Makhoul:"Segment Quantization for Very-Low-Rate Speech Coding", Proc. ICASSP, pp.1565-1568 (1982).

[5] J.Picone and G.R.Doddington:"A Phonetic Vocoder", Proc. ICASSP, pp.580-583 (1989).

[6] Frank K.Soong:"A Phonetically Labeled Acoustic Segment (PLAS) Approach to Speech Analysis-Synthesis", Proc. ICASSP, pp.584-587 (1989).

[7] S.Nakagawa:"Connected Spoken Word Recognition Algorithms by Constant Time Delay DP, O(n) DP and Augmented Continuous DP Matching", Information Sciences 33, pp.63-85 (1984).

[8] H.Ney:"The use of a One-stage Dynamic Programming Algorithm for Connected Word Recognition", IEEE Trans. Acoust., Speech & Signal Process., ASSP-32,2, pp.263-271 (1984)

[9] H.Sakoe:"Two-level DP-matching - a Dynamic Progrmming based Pattern Matching Algorithm for Connected Word Recognitions", IEEE Trans. Acoust., Speech & Signal Process., ASSP-27, 6, pp.588-595 (1979).

[10] S.Nakagawa and Y.Hashimoto:"A Method for Continuous Speech Segmentation using HMM", Proc. Int. Conf. Pattern Recognition, pp.960-962 (1988).