Pergamon

Expert Systems With Applications, Vol. 9, No. 1, pp. 81–89, 1995

0957-4174/95 $9.50 + .00

0957-4174(94)00051-4

Fuzzy Expert System for Continuous Speech Recognition

HA-JIN YU, YUNG HWAN OH

Department of Computer Science, Korea Advanced Institute of Science and Technology, Taejon, Korea

YOICHI YAMASHITA, RIICHIRO MIZOGUCHI

The Institute of Scientific and Industrial Research, Osaka University, Ibaraki, Japan

Abstract—*In designing a speech recognition expert system, there are a number of aspects to be considered—segmentation method, recognition unit, structure of the rule, and user aids. To avoid the difficulties in segmentation, an irregular unit based on spectral transition measure is proposed. Frames are used as the structure of the speech recognition rules. The structure provides the user with an easy way of making rules, and enables the building of an automatic rule generator. Fuzzy linguistic variables are used in representing rules. The rule-generating cycle, which includes a rule generator, a rule tester with an error reporter, and a state describer with other modules, could save much time in making rules and could result in high performance. The experimental results and comparison with the SPREX system, which is the basis of this system, are described.*

1. INTRODUCTION

SPEECH is one of the most common, effective methods of communication, and speech recognition has many fields of application. Numerous methods for speech recognition have been developed, such as template matching methods, statistical methods, neural network based approaches, and so on. However, most of the approaches have not yet been able to achieve the final goal of speech recognition, that is, speaker-independent continuous speech recognition. Most of the approaches still depend on heuristic knowledge in addition to their own methodology. If we can simulate the knowledge of spectrogram reading experts effectively, we will be able to get a recognition performance close to that of human experts. The main goal of our system is to provide the most effective way of extracting the knowledge of human experts and to achieve recognition performance as high as that of human experts.

A number of speech recognition expert systems have been proposed (Mizoguchi, Tsujino, & Kakusho, 1986; Stern, Eskenazi, & Memmi, 1986; Zue & Lamel, 1986). These systems provide powerful tools for transforming experts' knowledge into rules, and maintaining the rules effectively. However, they have the drawback that the labeling uses thresholds. The rules cannot be flexible for values near the threshold value and it is hard to determine exact threshold values. In this article, the concept of fuzziness is applied to the system to provide flexibility of rules, and to give more leeway to human experts who produce them. We use linguistic variables instead of thresholds to describe the state trajectory of parameters.

Rules are represented in frames in this system. In order to write a rule, one simply fills in the slots of the frames. This way of writing rules makes the automatic generation of the rule easy. In speech recognition systems, the segmentation problem is one of the difficult problems to solve. To overcome the problem, we defined a recognition unit based on a spectral transition measure so that the segmenter segments more easily. The system consists of the following modules: segmenter, automatic rule generator, rule tester and error reporter, state describer, and data bases for speech data and rules. The automatic rule generator provides the cue in making rules, and the error reporter and state describer with other modules in this system can save much time in making rules and improve the recognition performance.

2. UNDERLYING IDEAS

In this section, the concepts proposed in the new system are described after the brief review of the SPREX

Requests for reprints should be sent to Ha-jin Yu, Department of Computer Science, Korea Advanced Institute of Science and Technology, 373-1, Kusong-dong, Yusong-gu, Taejon 305-701, Korea.

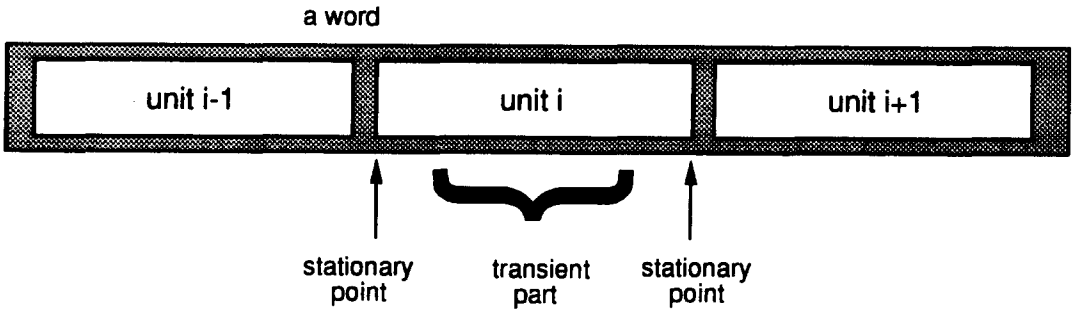

FIGURE 1. The definition of recognition unit.

system (Mizoguchi, Tanaka, Fukuda, Tsujino, & Kakusho, 1987), which is the basis of the new system.

2.1. Brief Review of SPREX

SPREX is a continuous speech recognition system using knowledge engineering techniques. Parameters are described in symbolic form, and then forward chaining rules are applied for recognition. The parameters used are formant frequencies, energy, zero crossing rate, and ratio of low frequency energy to the whole energy (L/A). Variations of the feature parameters are described in terms of descriptors such as rapid increase (I3), moderate increase (I1), constant (C), and so on. States of features are labeled by using the labels: N (noise), I (increase), D (decrease), C (constant), A (appearance), E (extinction), and G (gap). The state descriptions of feature parameters provide a good graphical interface for users who produce rules.

The rules can be written according to the knowledge of human experts. The key features used for writing the rules are not the absolute values of the parameters. Rather, the feature changes are important for recognizing the phonemes in various environments. For example, to position a dip in power, it is important where the power state changes D I (decrease and increase), and the exact value of the power in the point does not matter. The rules are stored in a rule data base and used for recognition. In the recognition process, the state description and the rules are used for segmentation, consonant recognition, vowel recognition, and rerecognition. Because it is difficult to realize discrimination among unvoiced plosives in continuous speech uttered by unspecified speakers, recognition categories are not divided into the respective phoneme themselves, but divided into phoneme groups. More detailed discrimination can be done by using contextual knowledge.

2.2. Recognition Units and Segmentation

In speech recognition systems, the segmentation problem still remains to be solved. The importance of the problem is that it is usually the first step in the recognition process. Errors in the segmentation step will propagate to later steps, and the performance of the total system can never rise above that of the segmenter. Moreover, because writing a rule is dependent on the segmented result, the rule should cover all cases even when a number of units are segmented as one.

We try to solve this problem by assuming that the segmenter is perfect, so that we can trust the results of the segmenter. We do not insist that the segmenter segment the input into the units we defined, such as phoneme, syllable, or else. Rather, we accept the result of segmentation as it is, and define it as the recognition unit of the system. The segmentation result may be a phoneme, sylla-

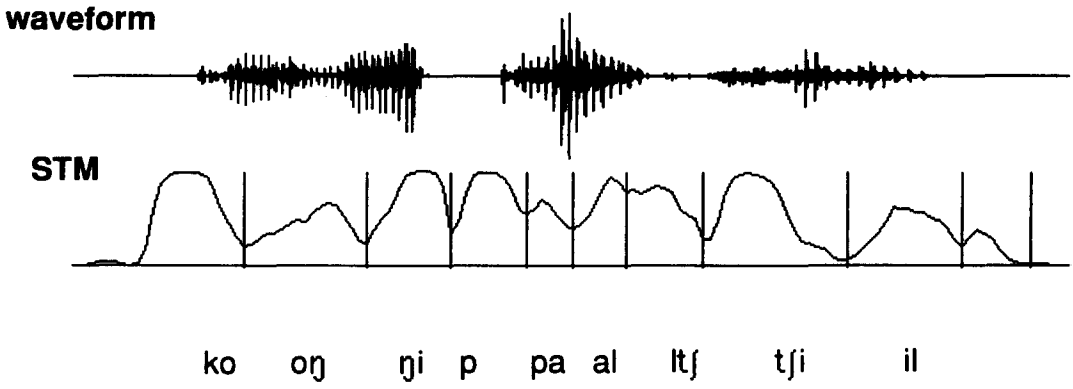

FIGURE 2. Segmentation using spectral transition measure (0287 in Korean).

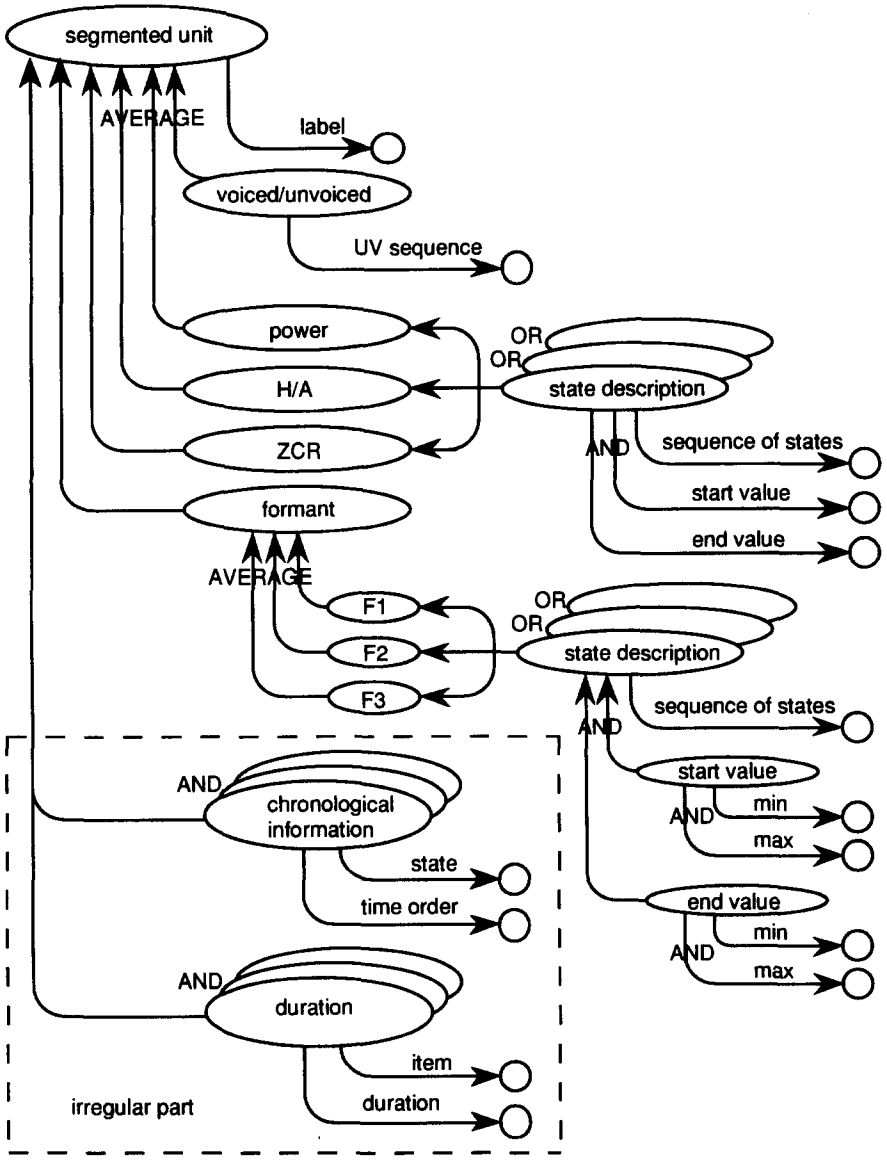

FIGURE 3. The rule structure.

ble, or word. The same words can be segmented into different units in various contexts. However, if we assume that the segmented result does not vary in the same environment, the segmented result can be defined as a unit.

The recognition unit is defined as follows. Every unit has its stationary point at each end of the unit as shown in Figure 1. A stationary point may be found mainly in vowels. Nasal sound, silence, and sometimes consonant can also include stationary points. The stationary point can be detected by spectral transition measure (STM) (Furui, 1986). The point where the value of STM is the local minimum is defined to be the stationary point. The reason for defining the minimum, not the maximum, of the STM to be the cutoff point is that, near the point where the value of STM is high, the spectral transition changes rapidly, so the segmentation should be very precise. Near the point where the change is slow, the small shift in cutoff point has little effect on the segmentation result. Figure 2 shows an example of STM used in the segmentation of continuous words.

As the result, the segmented unit can be a vowel-consonant-vowel (VCV) chain, a phoneme, a syllable, a word, or none of these. The point here is that there can be no error in the segmentation result because most of the segmented results are defined to be units. However, the units have the drawback that the number of them becomes too large. The large number of units makes the number of rules too large, and causes problems in making rules. These problems can be solved by automatic rule generation in later steps.

2.3. Structure of Rules

Frames are used as the structure of the rules. A frame is a collection of semantic net nodes and slots that

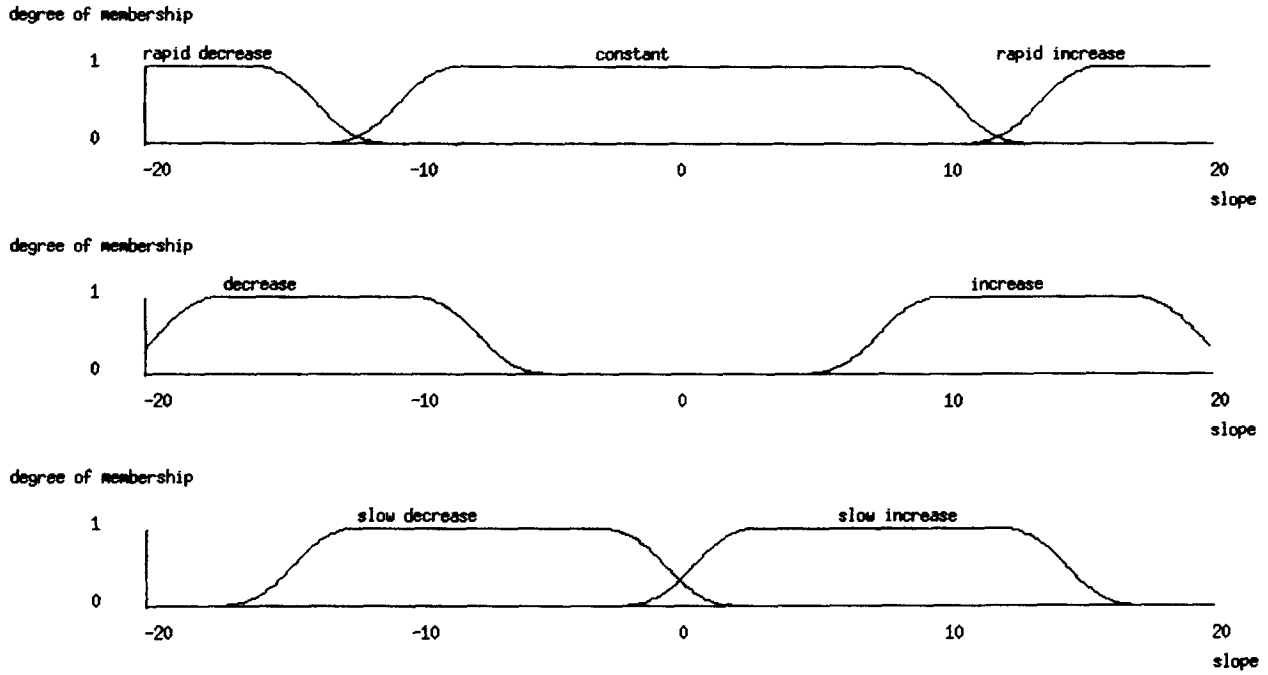

FIGURE 4. Membership functions describing the change of zero crossing rate.

together describe a stereotyped object. Figure 3 shows the definition of the rule structure. The parameters used in the rule are the same as those used in SPREX system. Each frame has slots for the state description such as the sequence of states, the start and end values, and so on.

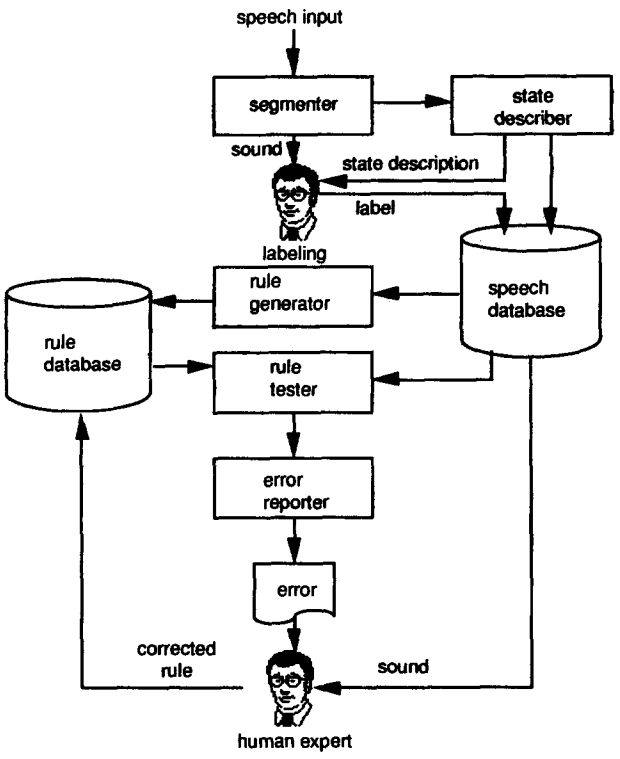

FIGURE 5. Flow of rule generation.

The rule is divided into two parts. One is the regular part, and the other is the irregular. The irregular part includes exceptional information such as that concerning chronology and duration. The regular part is filled by the rule generator and human experts, and the irregular part is filled by human experts only.

A node for a parameter in a frame has slots for values of the parameter and the change of the parameter. Only the start and end values of the parameter are used as the value of a parameter, because the absolute values in the unstable interval have little meaning. In the unstable interval, the changes of the values represent the unit, so the transition of the parameters are described as the state transition symbols. The stable points of the segments are the start and the end points from the definition of the unit.

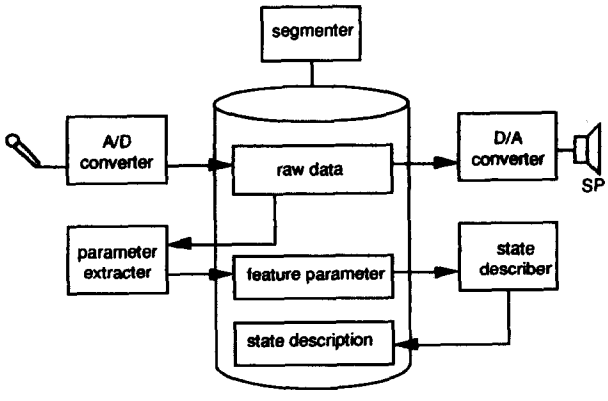

FIGURE 6. Speech data base.

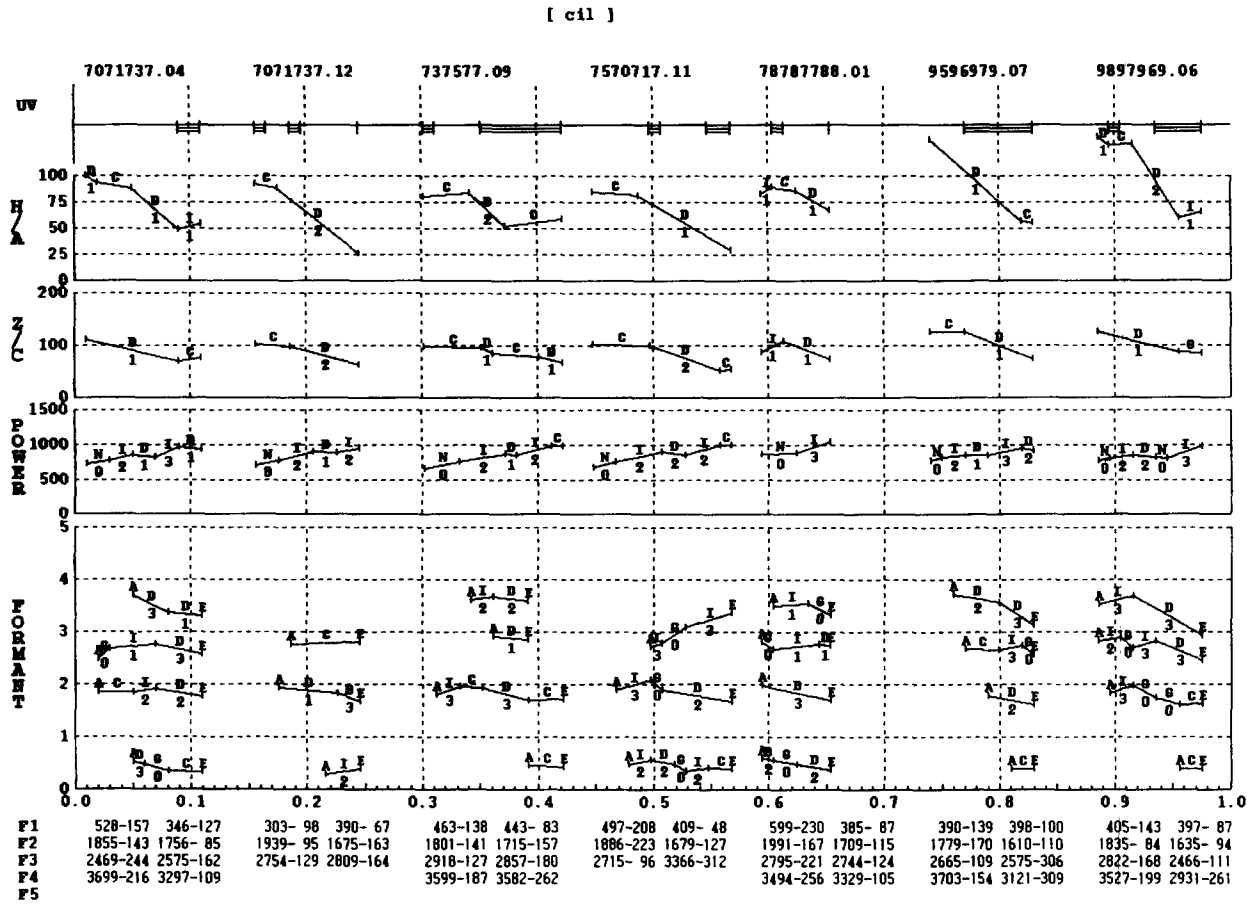

FIGURE 7. A state description for segments.

With this structure, writing a rule is simply filling in the slots of the frames. This way of writing rules makes their automatic generation very easy. There is no need to define a language for rules, and so there is no need to build a translator for translating rules made by humans into machine-recognizable form. Each value of the slots filled by humans can be used directly in the reasoning process.

2.4. Use of Fuzzy Variables

Most speech expert systems use qualitative descriptions, such as low, middle, and high, which are associated with disjointed numerical regions (Mizoguchi et al., 1987; Stern et al., 1986; Zue et al., 1986). This strict distinction of the levels is inflexible in the case that a value is slightly over a threshold value. An expert will feel free if he/she can describe the state change of parameters in natural language. In our system, the states of the parameters are described by linguistic variables such as rapid increase or slow increase. No numerical values except the formant frequencies are filled into the slots.

For the membership functions of the linguistic variables, the standard functions with adjustable parameters are used (Zadeh, 1987). The standard functions can be defined as Equation 1.

$$
\begin{aligned}
S(v;\alpha,\beta,\gamma) &= 0 \quad \text{for} \quad v \le \alpha \\
&= 2\left(\frac{v-\alpha}{\gamma-\alpha}\right)^2 \quad \text{for} \quad \alpha \le v \le \beta \\
&= 1 - 2\left(\frac{v-\gamma}{\gamma-\alpha}\right)^2 \quad \text{for} \quad \beta \le v \le \gamma \\
&= 1 \quad \text{for} \quad v \le \gamma
\end{aligned} \tag{1}
$$

In $S(v;\alpha,\beta,\gamma)$, the parameter β is the crossover

TABLE 1
Analysis Conditions of Input Speech

sampling rate	10 KHz
frame shift	10 msec
window	20 msec Hamming
number of speakers	10 males
total number of words	4 digits × 35 × 10 speakers × 4 times = 5600

Defined in rule generation

a	ach	ag	agu	ai	ail	ailo	al	alch
alchi	als	am	ami	ao	ar	aryu	as	al
ch	chi	chil	g	gi	go	gong	gu	gui
gus	i	ich	ig	igo	iil	iis	il	ilch
ilg	ili	ilo	ils	io	is	isa	iyuk	k
ko	kong	ks	ku	l	lch	lchi	lg	li
lo	lryu	lryuk	ls	m	mch	mgo	mi	mil
mgu	mn	mo	ng	ngch	nggo	ngi	ngil	ngn
ngnyu	ngnyuk	ngo	ngs	nyu	nyuk	o	och	ochi
og	ogu	oi	oil	ong	ongch	onggo	ongi	ongs
oo	os	p	pa	pal	pali	ryu	ryuk	s
sa	sam	samy	u	uch	ugi	ui	uk	uo
uos	ur	uryu	us	y	yu	yuk	yuks	yung

Appeared in test data

ms	moi	uks	yugo

FIGURE 8. Recognition units.

point, that is, the value of v at which S takes the value 0.5. The linguistic variable "slope" is depicted in Figure 4.

3. STRUCTURE OF THE SYSTEM

Figure 5 shows the overall structure of the speech recognition expert system, and the flow of rule generation. In this section, the sequence of rule generation is described.

3.1. Segmenter

Input speech is segmented into various units by the STM as explained above, and stored into the speech data base.

3.2. Labeling by Humans

The segmented units are labeled by human experts. Two kinds of information are given to the expert—the sound and the state transition diagrams.

3.3 Speech Data Base

The segmented raw data are stored in the speech data base. All of the parameters from the segmented data are extracted by the parameter extractor, and the parameters are saved into the speech data base. The state descriptions are also stored in this speech data base. Figure 6 shows the structure of the speech data base and the relations with other modules.

state transition	number of occurences	file name of the data
D1	37	d1-12-1.1 d1-14-2.6 d1-15-1.6 d1-15-2.5
D1 C	25	d1-12-3.1 d1-15-1.8 d1-15-2.7 d1-15-3.7
D1 I1	7	d2-1-2.6 d2-1-4.6 d2-12-1.1 d2-15-1.7
I1 D1	4	d1-12-2.2 d10-1-4.8 d10-2-4.3 d4-1-1.7
C D2	2	d1-1-1.7 d2-2-3.2
D1 C D2	2	d1-2-2.1 d8-12-2.1
D1 C D3 C	2	d2-15-1.5 d9-2-3.3
C D3 C	2	d4-2-4.3 d7-1-2.8
D1 C D1	2	d8-12-1.1 d8-14-3.5

FIGURE 9. An example of statistical information (zero crossing rate for unit "chil").

```
RULE chil-1

UV
        U  V
PWR
        high low
        N  I3                OR
        N  I3 D1
HA
        high low
        D1 C                 OR
        D1                   OR
        C  D1                OR
        C  D1 C
ZC
        high low
        D1                   OR
        D1 C                 OR
        D1 I1                OR
        I1 D1
F1
        D                    OR
        C
        sv  395-1028
        ev  360- 910
F2
        sv 1338-2209
        ev 1586-2510
F3
        sv 1749-3353
        ev 1754-3435

CHRONOLOGY

        HA D.E < PW I.S   AND
        ZC D.E < PW I.S

DURATION

        PW I    < ZC D    AND
        PW N    < ZC D
```

FIGURE 10. An example of recognition rule.

3.4. Rule Generator

The rule generator produces temporal rules automatically, and stores them in the rule data base. Segments of the same label are gathered together to acquire statistical information, such as the most probable state transition sequences. Using this information, the slots in the frames of the rule are filled. The frequent state transition sequences are entered into the slots for state description, and the sequences are related by OR operation. For the formant values, only the maximum and minimum values of the start and end points are filled into the slots. For the values of other parameters, the values are transformed into linguistic words, and only the words are filled into the slots. The slots for the irregular part are not filled by the rule generator because the amount of time ordering information of all of the state transitions is too large and unnecessary. These slots are filled later by the human expert, if needed.

3.5. Rule Data Base

The rules generated by the rule generator and human experts are stored in the rule data base. The rules are divided into three groups. The first one consists of the rules for general speakers, the second for each particular speaker, and the rest for groups.

3.6. Rule Tester

All of the segments of test data set with labels are tested by the rules in the rule data base. Each input slot is compared with each slot of all the rules. For a state description, the tester sees if the state transition sequence in rule slots are in the test slots. Later, the degree of membership is calculated for the slope of each state. The degree of membership of the states in the state sequence is summed up by AND operation as shown in Figure 3. In the case that no exact matching of the state sequence is found, a state sequences in a rule slot, which is included in the test slot, in due order is selected. Finally, the degree of membership for a parameter in a rule is summed up by the operations shown in Figure 3.

Maximum operation is used for OR operation, and minimum operation is used for AND operation. The degree of membership of the start and end values of each parameter is also calculated. The final score is calculated by averaging the degrees of membership of all the parameters. The reason why the average was used is that one or more of the parameters, such as formant frequency, may not be present in some cases, leaving the other parameters to make up the properties of the unit. If MIN operation is used, the overall score becomes zero, and other important parameters can have no effect on the score.

3.7. Error Reporter

The error reporter shows the detailed steps of the recognition process. For each feature parameter of each tested unit segment, each slot of all of the rules and the matching scores in case of error are reported. So,

```
* Parameter file name      : d10-1-1.2

RULE name :    ongi

>>> ZCR
Rule : increase slow, low, low
Input :  C    78.0 -> 76.0   slope : -0.50
score : 0.18
# ZCR :  0.18

>>> F1
Rule  : sv 360-620  ev 430-750
Input : sv 364  ev 405
score : start value  1.00
score : end  value  0.69
# F1 :  0.69

# ongi :  0.86

INPUT : ongi
RESULT : mi
```

FIGURE 11. An example of an error report.

it is very easy for the user to find out where the error is in the rule, and to find out how to modify the rule.

3.8 State Describer

The state describer shows the state description of different segments of the same unit in one panel, so that the user can grasp the common property of the unit at a glance. It frees the user from having to keep swapping from one sentence to other sentences to find out the common property of a unit in many sentences. In fact, it may be a difficult job for the human who produces the rules to search a lot of sentences. The automatic rule generator, error reporter, and state describer can save much time for making rules. Figure 7 shows an example of the state description. The figure shows seven segments of a unit "chil," which means seven in Korean. In the figure, the top of the panel shows the name of the unit, and located below the name is the name of the files where each segment comes from. UV section shows the voice/unvoice state of each unit. The twofold line segments show where there are voiced parts in the unit. Since the unit begins with unvoiced consonant, the beginning of each segment is marked as "unvoiced." The rear parts of the segments are marked as "voiced." The ratio of high frequency energy, H/A, and zero crossing rate, Z/C, show decreasing characteristics, since the values of those of consonants are high, and those of vowels are low. The energy of the segment is increasing in time as shown in POWER section. The state transitions of formant frequencies and the range of their values are shown in the lower part.

3.9. Refinement of the Rule by Human Effort

The human expert can find out where the error exists in the rules with ease using the error report and can correct the error. A further job for the human will be adding the chronological and durational information of each parameter, so that the increasing state of power appears after the decreasing state of zero crossing rate, etc. The human expert is allowed to add a intermediate value of a parameter if it is necessary. He can also listen to the segment whenever he wishes. The segment can be relabeled in this phase. The rules corrected by humans are fed back to rule tester, and the testing cycle continued until the recognition rate becomes satisfactory.

4. EXPERIMENTS AND RESULTS

The system was tested by speech data of four connected Korean digits. There are 35 digit sequences in the data set, and each digit sequence was uttered four times by ten male speakers. Table 1 shows the analysis conditions of the data. The data uttered by five speakers are used in the rule generation, and the rest of the data are used in the rule test. Figure 8 shows the segmented result of a digit sequence, and the units defined. In rule testing with new data, four kinds of extra segments appeared. The result of segmentation is dependent on speakers, so new units were defined additionally in the test. More units may have to be added to adapt to new speakers in the future.

Figure 9 shows state transition patterns of a unit produced by the rule generator. The second column is the number of the segmented data of that pattern, and the third column shows the name of the data. Figure 10 shows an example of a rule, and Figure 11 shows a small part of an error report. Figure 12 shows the recognition rate versus the number of cycles. The recognition rate of the first cycle in the graph is the result of the rule tester with the rules generated by the rule

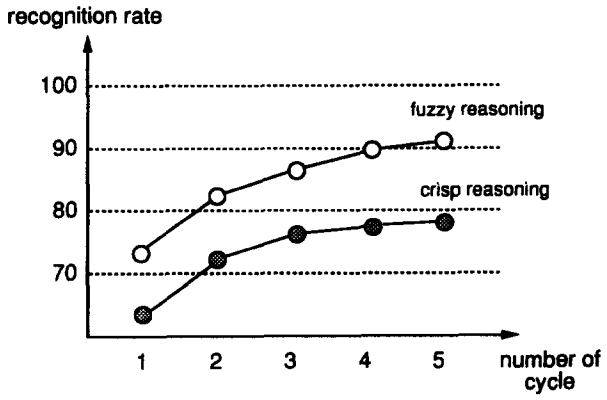

FIGURE 12. Recognition result for each rule generating cycle.

TABLE 2
Comparison of SPREX II and the Proposed System

	SPREX II	Improved SPREX
Rule structure	RL/SR	frame
Recognition unit	phone group	irregular unit based on STM
Rule variable	numeric	linguistic
Distinction of levels	threshold	fuzzy membership
Rule test	by human	by rule tester and error reporter
Listening capability	not provided	provided in rule generation phase
Target machine	workstation and symbolics	workstation

RL/SR = a rule language for speech recognition.

generator, and the values following are the result of each cycle with rules corrected by humans.

The figure also shows the comparison of the two reasoning method, the crisp reasoning and the fuzzy reasoning. We obtained the result of the crisp reasoning by replacing the memberships function with thresholds. It shows that the fuzzy reasoning reduced the error rate by 58% in the last cycle. The recognition rate is currently 91.2%, but the result shows that the performance of the system can rise with each cycle. The time required to make a rule is also reduced to about $\frac{1}{5}$ of that required without the rule generator.

5. CONCLUSION

A continuous speech recognition system using knowledge engineering techniques and fuzzy concepts has been discussed. The main goals of the system are to provide the most effective way of extracting the knowledge of human experts and to achieve a recognition performance as high as that of human experts. The system provides the user with a graphical and natural interface by means of the state description and linguistic variables. Table 2 shows the configuration of the system with comparison with the SPREX II system.

A new recognition unit is defined to reduce the error rate in segmentation. The segmentation is based on the stationary parts where the value of the spectral transition measure is low. Frames are used for the structure of the rules, which provides the user with an easy way of making rules, and makes it easy to build an automatic rule generator. As the experimental result shows, the rule-generating cycle that consists of rule generator, rule tester with the error reporter, and state describer saved much time in making rules and improved recognition performance. It is shown that using fuzzy membership improves the recognition performance, and the recognition unit also turned out to be effective.

We expect that the performance of the system can be improved constantly by updating the rules with the rule generating cycle. Contextual knowledge might be effective for recognizing words and sentences.

Acknowledgment—The authors would like to thank the anonymous referees for their helpful comments, and appreciate the support to international cooperative project given by Korea Science and Engineering Foundation. The first author is also grateful to Kazuhiro Arai and Seong-jin Yun for help in programming.

REFERENCES

Furui, S. (1986). On the role of spectral transition for speech perception. *Journal of Acoustic Society of America* **80**(4), 1016–1025.

Mizoguchi, R., Tsujino, K., & Kakusho, O. (1986, April). A continuous speech recognition system based on knowledge engineering techniques. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing '86* (pp. 1221–1224), Tokyo, Japan.

Mizoguchi, R., Tanaka, Y., Fukuda, N., Tsujino, K., & Kakusho, O. (1987). A continuous speech recognition expert system—SPREX. *The Transactions of the Institute of Electronics, Information and Communication Engineers,* **J70-D**(6), 1189–1198.

Stern, P., Eskenazi, M., & Memmi, D.(1986, April). An expert system for speech spectrogram reading. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing '86* (pp. 1193–1196), Tokyo, Japan.

Zadeh, L. A. (1987). A fuzzy-algorithmic approach to the definition of complex or imprecise concepts. *Fuzzy sets and applications: Selected papers by L. A. Zadeh.* Philadelphia, PA: John Wiley & Sons, Inc.

Zue, V., & Lamel, L. (1986, April). An expert spectrogram reader: A knowledge-based approach to speech recognition. *Proceedings of International Conference on Acoustics, Speech, and Signal Processing '86* (pp. 1197–1200), Tokyo, Japan.