Pitch Pattern Clustering of User Utterances in Human-Machine Dialogue

Takashi YOSHIMURA, Satoru HAYAMIZU, Hiroshi OHMURA and Kazuyo TANAKA

Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba, Ibaraki 305, JAPAN
E-mail: yoshimur@etl.go.jp

ABSTRACT

This paper argues about pitch pattern variations of user utterances in human-machine dialogue. For intelligent human-machine communication, it is essential that machines understand prosodic characteristics which imply a user's various attitude, emotion and intention beyond vocabulary. Our original focus is on particularly distinct pitch patterns and their roles in the actual dialogues. We used human-machine dialogues collected by a Wizard of OZ simulation, Many utterance segments belonged to clusters that were prosodically flat patterns. From the result, we considered that utterances which belonged to the other clusters and those which were far from the centroids included non-verbal information. In these utterances, there were talks to themselves and questions to the machine including emotional expressions of a puzzle or a surprise. These pitch patterns were not only rich in ups and downs, but also their slopes were upward, while the pitch pattern were generally even or a little downward. These results indicate that peculiar pitch period patterns show non-verbal expressions. In order to actually utilize such information on human-machine interactions, the representative pitch patterns should be investigated concerning their relationship to various types of communication.

1. INTRODUCTION

In human-human communication by speech, we emphasize a essential point in dialogue, turn a topic and express various attitudes, emotions and intentions, which are not able to be transmitted with vocabulary, by changes of prosodic characteristics. As well as for intelligent human-machine communication, it is essential that machines understand some prosodic characteristics and also generate some prosodic characteristics which satisfy a demand in each scene[1]. Past works for prosodic analysis, however, mainly dealt with human-human imitated dialogues or read sentences, that is not human-machine communication data. They detected phrase boundaries[2] and estimated the sentence structure[3].

We analyze actual human-machine communication data, and examine prosodic features. Prosodic features has several aspects[4]: accent features[5], intonation features and so on. In this paper, we focus on pitch pattern variations of the user's utterances in human-machine dialogues. We suppose that when users want to convey information beyond expression, they use specific pitch patterns which are different from common pitch patterns. So we classify pitch patterns of user utterances and argue the relation with verbal information about particularly distinct pitch patterns from others.

First, in the next section, we describe dialogue data we used. In the third section, we discuss in detail our method used for pitch pattern clustering of user utterances. We show the experiments of pitch pattern clustering in the fourth section. The relationship of expressions about some peculiar utterances which have different pitch patterns are also discussed.

2. HUMAN-MACHINE DIALOGUE DATA

2.1. Data collection

In this research, we used human-machine dialogues[6]. The data were collected by a Wizard of OZ simulation. The task was related to town information of Shibuya in Tokyo. In this dialogue data, user utterances and synthesized machine utterances were sepalatedly recorded to different recording channels at the same time. Only user utterances were analyzed in this work.

2.2. Speech segments

This sample set contained 7,495 utterances units pronounced by 40 speakers (males and females), both which were detected as a speech segment and where pitch pattern were extracted. These units were detected using power and zero-crossing values of speech wave forms. There were considered to segment boundaries when silences continued more than 300 msec. We supposed that one speech segment corresponds to one utterance unit.

Table 1 shows the length distribution of utterance units. The result indicates that most utterances were less than 2 sec.

length	number
less than 1 sec.	4154
1 sec. to 2 sec.	2328
2 sec. to 3 sec.	768
3 sec. to 4 sec.	179
4 sec. to 5 sec.	39
more than 5 sec.	27
Total	7495

Table 1: The length distribution of utterance units

pitch pattern cluster	number
No. 0	1274
No. 1	578
No. 2	2899
No. 3	547
No. 4	1034
No. 5	298
No. 6	630
No. 7	235
Total	7495

Table 2: Number of utterance units in each pitch pattern cluster (VQ size: 8)

3. PITCH PATTERN CLUSTERING

As automatic pitch extraction, we used a fundamental wave filtering method which is especially effective to fine pitch pattern extraction[7]. In former methods, avarage pitch were mainly obtained in every moving observation window whose width and period were previously defined. In this method, we obtain a pitch period that is equivalent to the length of a fundamental wave. Representative pitch patterns of utterance units were calculated as follows:

(1) Log pitch periods were warped to 32 sample points for each unit. If the difference between a sample pitch period and the interpolated value of adjoining pitch periods was over the threshold, the interpolated period was substituted for the sample period.

(2) The first regression coefficients and normalized residuals of these smoothed pitch period patterns were calculated. The normalized residual pitch patterns were used for making representative pitch patterns. The slope data were for observation of prosodic characteristics.

(3) The normalized residual pitch patterns were regarded as 32 dimensional vectors, and then quantized by vector quantization. Each centroid was the representative pitch pattern.

4. CLUSTERING EXPERIMENTS

Using human-machine dialogue data, as described in Section 2.1 and 3, representative pitch patterns were automatically generated from several VQ codebooks, whose sizes were 4, 8 and 16. Fig. 1 and Fig. 2 show representative pitch patterns whose codebook size is 8. Table 2 also shows number of utterance units in each pitch pattern cluster whose codebook size is 8.

The results indicate that many pitch patterns belong to cluster No.0 and No.2, so many utterrance segments are prosodically flat patterns. From the results, we considered that utterances which belonged to other clusters and those which were far from the centroids included non-verbal information. We confirmed their transcriptions and checked by listening. In these utterances, there were talks to themselves and questions to the machine including emotional expressions of a puzzle or a surprise. These pitch patterns were rich in ups and downs.

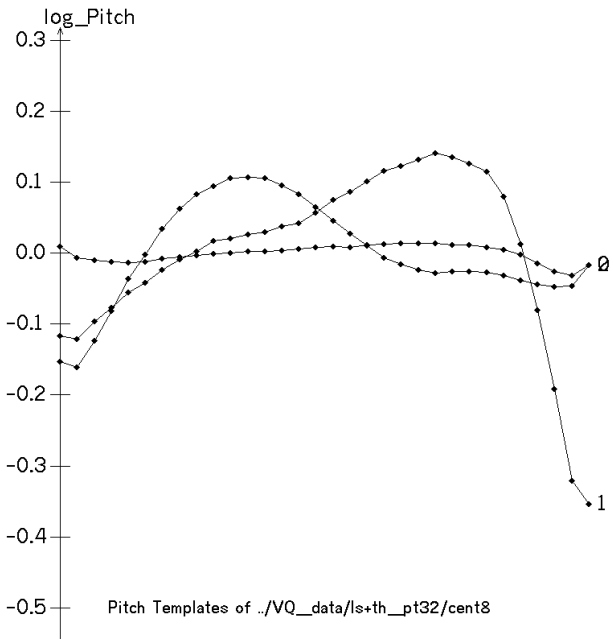


Figure 1: Representative pitch patterns (VQ size: 8, No.0 to No.2)

Table 3 and Table 4 show example transcriptions including utterances of peculiar pitch patterns. Fig 3 and Fig 4 also show the pitch patterns of example 1. and 2. These two utterances belong to cluster No.6 and each pitch pattern rises and falls two times. In the underlined utterance of example 1., we confirmed that changes of the pitch pattern just after laughing are much bigger than other common declarative utterances. In the underlined utterance of example 2., we checked that vowels continue longer at the part of "...shi, de..." by listening. The figure also tells us the situ-

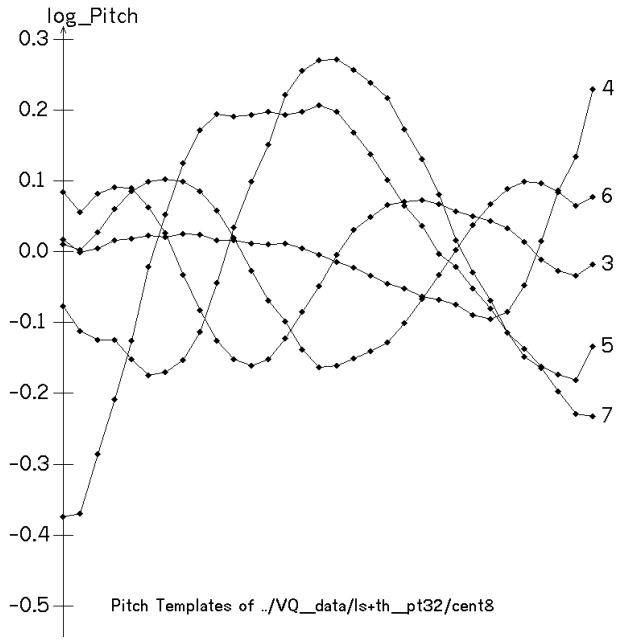


Figure 2: Representative pitch patterns (VQ size: 8, No.3 to No.7)

ation that the user is considering and talking at the part of "...shi, de...".

Table 5 shows the avarage slope distribution of pitch patterns by the linear regression analysis. The results indicate that the slope value of most utterances are from -0.02 to 0.01. So common utterance units have flat or slightly downward slopes. In few utterances whose slopes were sharp increases, there were utterances including emotinal expressions which were the same as former examples. They were different from the general mode of expression. The result indicates that the slope information is possible to be elements for non-verbal information extraction.

In few segments whose pitch patterns were far from any representative pitch patterns or were upward about the slopes, however, synthesized machine utterances were accidentally recorded together. In the future, the segments should be excluded from the objects of analysis and the user utterance set should be reclassified.

5. CONCLUDING REMARKS

For investigation of prosodic characteristics which are essential to human-machine communication, we analyzed human-machine dialogue data. From the result of pitch pattern clustering, most pitch patterns of user utterances are prosodically flat and do not include information beyond expression. So we examined some peculiar pitch period patterns in the data, which were far from the centroids of clusters. These

User: [Nto] hachikoomaeno koosateNkara,
dooyatteikuka wakaNnai
([well] I don't know the way from the crossroad
in front of the figure of Hachi)
System: hachikoomaeno koosateNyori
migikata, eeto
(to the right of crossroads
in front of the figure of Hachi, well)
seNroto heikooninobirumichio susumuto
(walk down the road along a railway and)
shiNgoonoarukoosateNga arimasu
(you will find crossroads with a signal)
User: seNrotoheekoo(nomi,)no michi^.
(seNrotoheekoo)
((the ro,) the road along a railway ?
(along a railway))
System: motto kaNtaNna shitsumoNni shitekudasai
(please phrase that more simply)
User: (warai) yokuyuuna.
((laugh) you say such a thing.)

Table 3: Example utterances 1.; an underlined utterance has a peculiar pitch pattern

results indicate that the pitch patterns are full of ups and downs and show some non-verbal expressions. In the view point of avarage slope by the linear regression analysis, the pitch patterns are also different from common pitch patterns which are generally flat or slightly downward.

In order to actually utilize such information on human-machine interactions, the representative pitch patterns should be investigated about relations to various types of communication. We also try to reexamine peculiar pitch patterns about the role in the actual dialogues and to extract para-linguistic information through the principal element analysis of pitch patterns.

6. ACKNOWLEDGEMENTS

The authors are indebted to Dr.N.Otsu, Director of the Machine Understanding Division in ETL, for his continued valuable support, and to the members of the Speech Processing Section for their useful discussions and technical assistance.

7. REFERENCES

1. S.Hayamizu: "Lively Communications with Spoken Dialogue Systems Utilizing Acoustic-Prosodic Information" LIMSI Technical Report, No.94-26 (1994-12)

2. M.Nakai and H.Shimodaira: "Accent Phrase Segmentation by Finding N-Best Sequences of Pitch Pattern Templates" Proc. ICSLP-94, **S08-10**, pp.347-350 (1994-9)

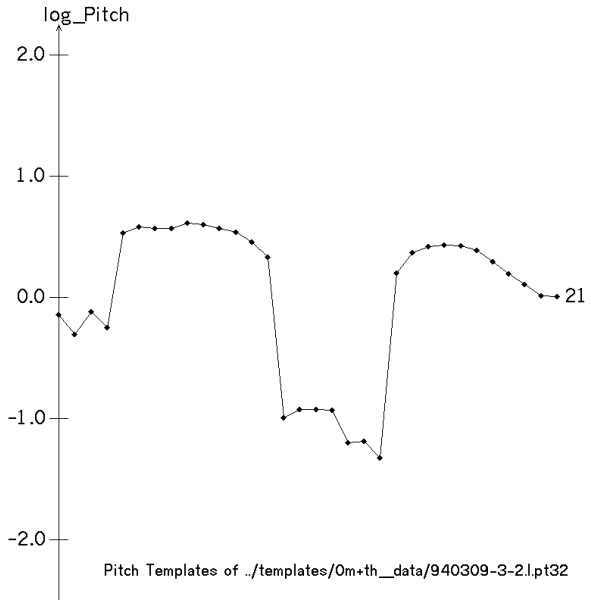


Figure 3: An example of peculiar pitch pattern; utterance unit 1. (**User:** `(warai) yokuyuuna.`)

3. A.Imiya, T.Sekiya and A.Ichikawa: "Estimation of Sentence Structure by Intonation" JSAI Technical Report, SIG-SLUD-9501-2 (1995-6) *in Japanese*

4. C.W.Wightman and M.Ostendorf: "Automatic Labeling of Prosodic Patterns" IEEE Trans. Speech and Audio Processing, Vol.2, No.4, pp.469-481 (1994-10)

5. T.Yoshimura, S.Hayamizu and K.Tanaka: "Word Accent Pattern Modelling by Concatenation of Mora Hidden Markov Models" Proc. IEEE ICASSP-94, **11.10**, pp.(I)69-(I)72 (1994-4)

6. K.Itou, T.Akiba, O.Hasegawa, S.Hayamizu and K.Tanaka: "Collecting and Analyzing Nonverbal Elements for Maintenance of Dialog Using a Wizard of OZ Simulation" Proc. ICSLP-94, **S17-10**, pp.907-910 (1994-9)

7. H.Ohmura: "Fine Pitch Contour Extraction by Voice Fundamental Wave Filtering Method" Proc. IEEE ICASSP-94, **76.6**, pp.(II)189-(II)192 (1994-4)

User: `wakarimashitakaˆ`
(do you make a sense ?)
System: `daitai, wakarimashita`
(generally, I can understand)
User: `korede, iidesukaˆ`
(now, are you alright ?)
System: `tochuuninanika, mejirushiwa arimasuka`
(on the way, are there any landmarks ?)
User: `mejirushi, desukaˆ.`
(you mean landmarks ?)
User: `omoiukabimaseN`
(I can't come to mind)

Table 4: Example utterances 2.; an underlined utterance has a peculiar pitch pattern

slope (log_pitch/point)			number
	less than -0.02		109
-0.02	to	-0.01	894
-0.01	to	0.00	3607
0.00	to	0.01	2601
0.01	to	0.02	201
more than 0.02			83
	Total		7495

Table 5: The avagage slope distribution of pitch patterns of utterance units

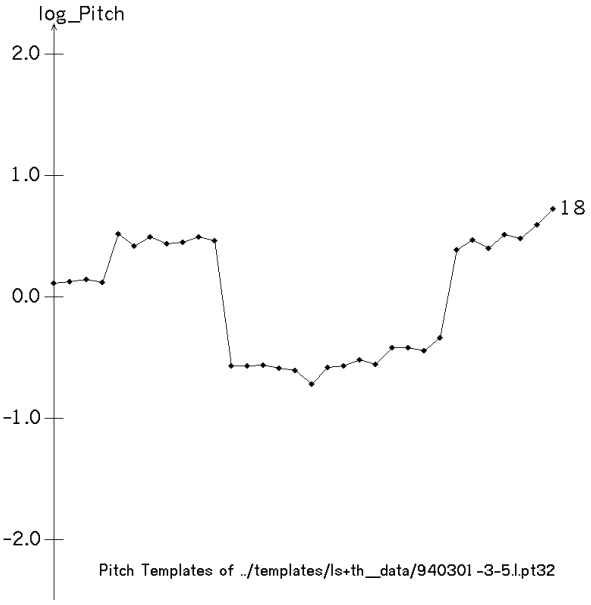
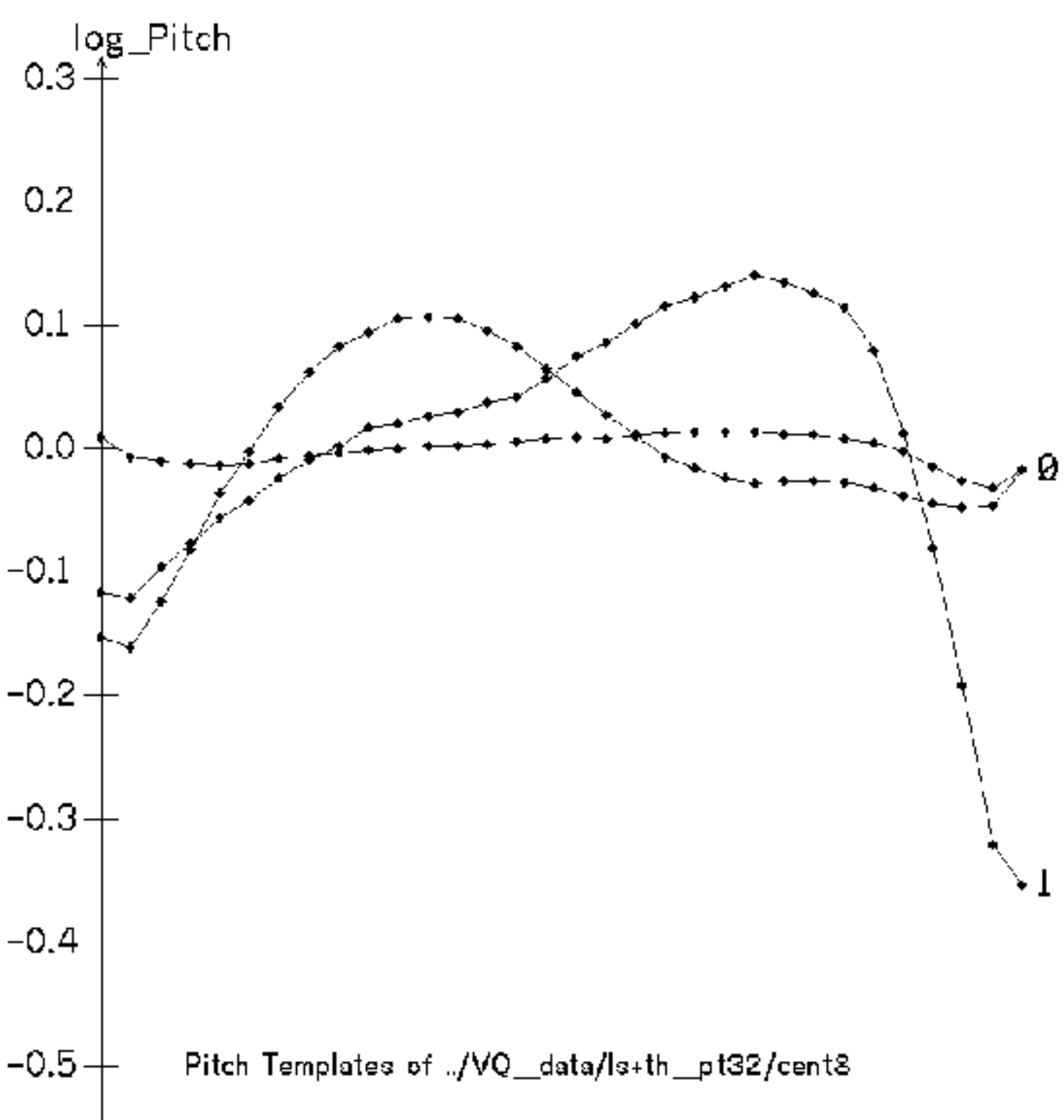
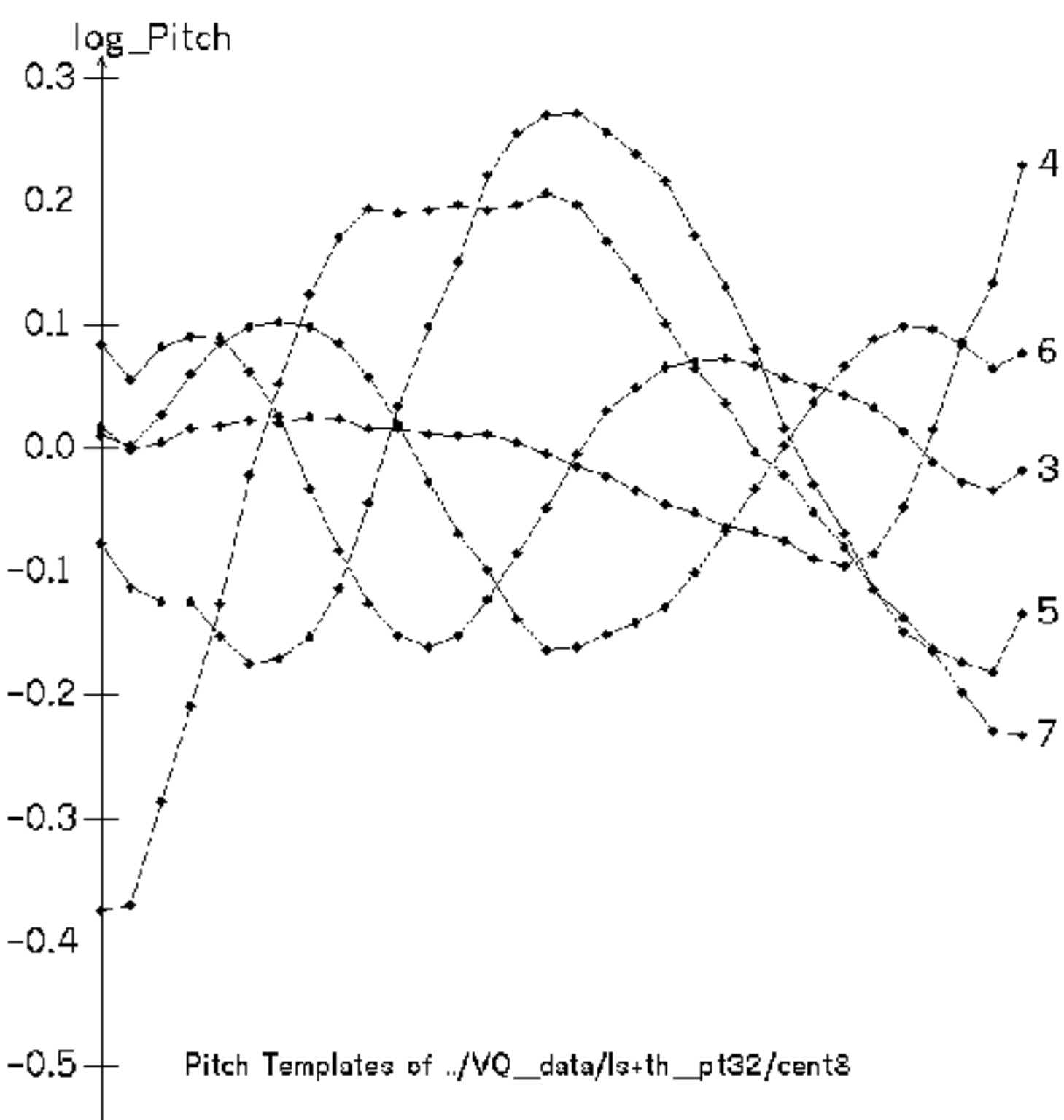
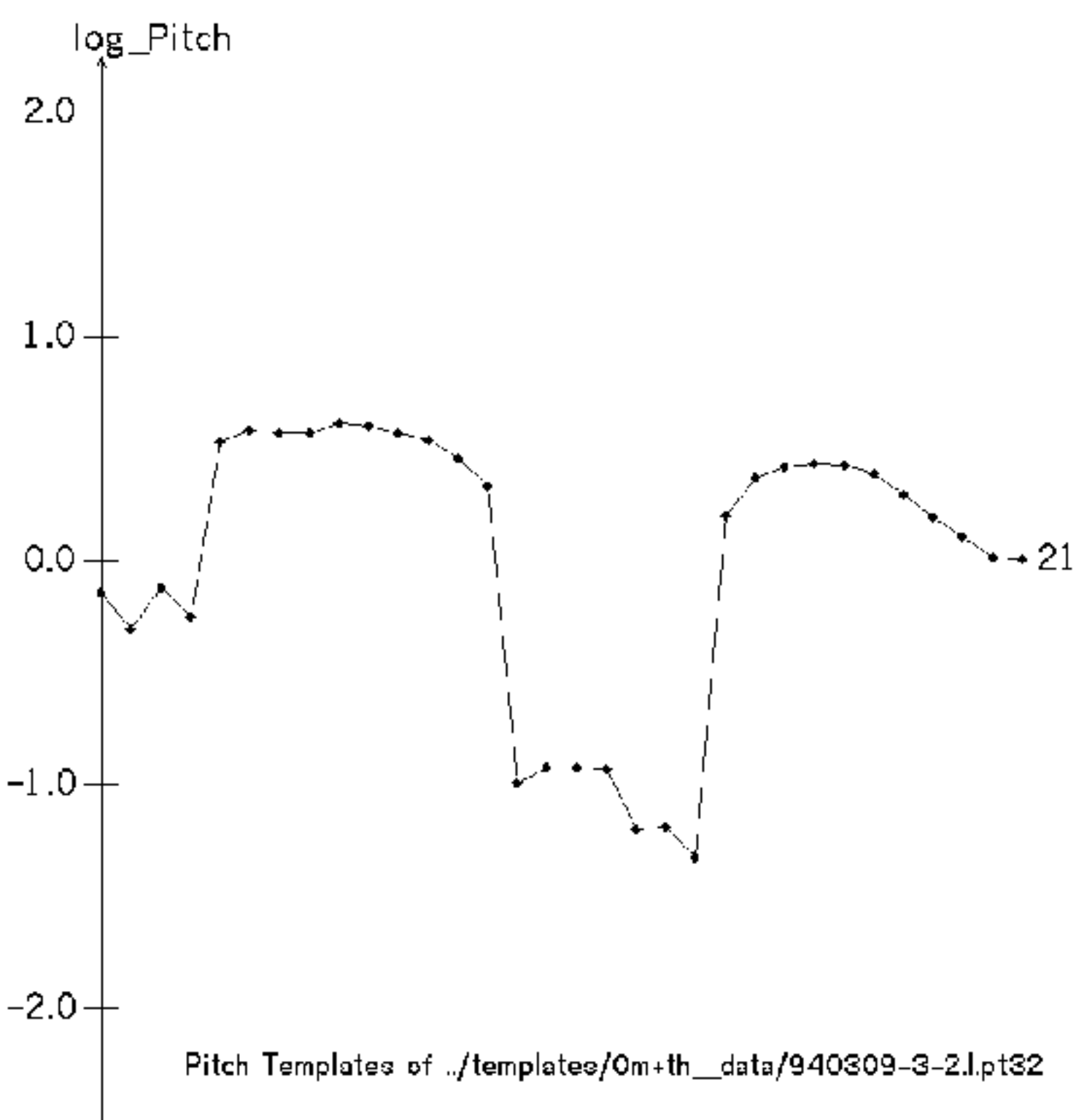
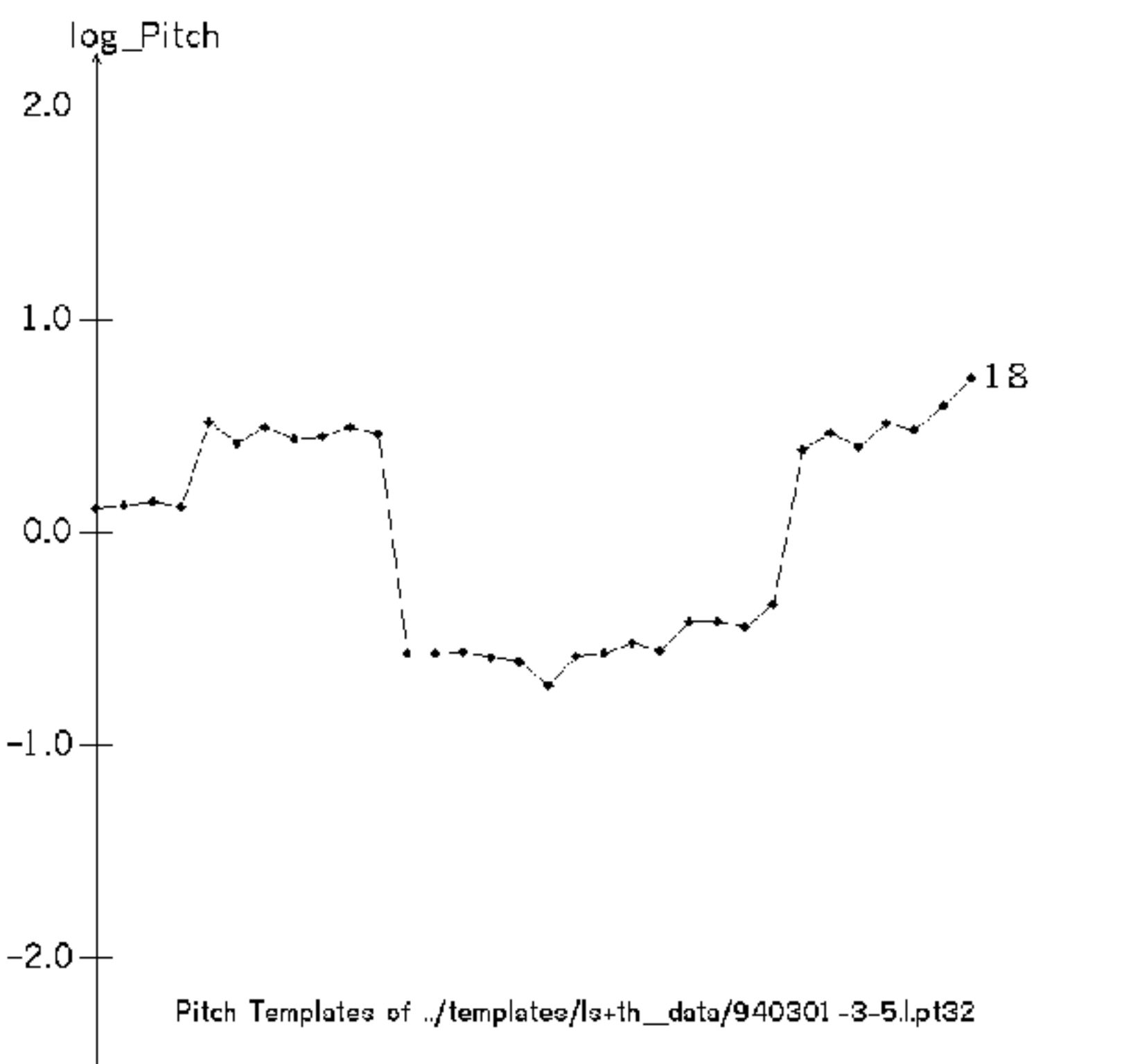


Figure 4: An example of peculiar pitch pattern; utterance unit 2. (**User:** `mejirushi, desukaˆ.`)


log_Pitch
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
0
1
Pitch Templates of ../VQ_data/ls+th_pt32/cent8


log_Pitch
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
4
6
3
5
7
Pitch Templates of ../VQ_data/ls+th_pt32/cent8


log_Pitch
2.0
1.0
0.0
-1.0
-2.0
21
Pitch Templates of ../templates/0m+th_data/940309-3-2.l.pt32


log_Pitch
2.0
1.0
0.0
-1.0
-2.0
18
Pitch Templates of ../templates/ls+th_data/940301 -3-5.l.pt32

Sound File References:

[a415_1.wav]

[a415_2.wav]