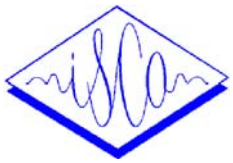

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Can People Perceive Different Emotions from a Non-emotional Voice by Modifying its F0 and Duration?

Yasuko Nagasaki & Takanori Komatsu

Department of Media Architecture
Future University - Hakodate, Japan
{yasuko; komatsu}@fun.ac.jp

Abstract

Forty-four stimuli were made from the unemotional utterance "eh" with duration changes (4 levels) and range of F0 (11 levels). Ten adult participants were asked to judge if the stimuli were congruent with the contexts (disagreement, hesitation, and agreement). Stimuli with rising tones tended to be identified as "surprise." On the other hand, stimuli with falling tones were identified as "postponement" when their duration was long, and were identified as "affirmation" when their duration was short. The results indicated that the duration and the ranges of F0 should be effective in identifying the contexts in which they were spoken.

10.21437/SpeechProsody.2004-154

1. Introduction

Much information can be extracted from spoken language: not only segmental but also super-segmental information, such as prosodic information [1]. However researchers in communication have been mainly focusing on segmental information [2]. Research is behind on super-segmental information such as pitch of the voice, duration, intensity, or voice quality. It may be because such super-segmental information of utterance was generally congruent with the utterance's literal meaning. In addition, there are only several studies with Japanese [3,4,5,6,7].

Communication between humans and machines may possibly become a reality, because of the development of speech recognition techniques. Thus, we cannot ignore super-segmental information, anymore. For example, the Japanese interjectory words "eh," "ah," or "ha" can be used in situations of agreement, disagreement, or hesitation to show one's thought. To make a machine which can guess the speakers' thoughts or their emotional states, an accumulation of super-segmental information would be needed.

The phoneme in "eh," "ah", or "oh," and other such interjectory words, used as a reply, seem to have a special meaning, similar to a baby's cooing. The phoneme will be selected not because the phoneme is suitable for the context. The speaker may pronounce the phoneme unconsciously. If such utterances are strongly influenced by the speakers' physiological changes along with his emotional state, the results may be applied to people whose native language isn't Japanese.

To start with testing Japanese listeners, we chose the word "eh" as a target for the sequential experiments. In Japanese, the interjection "eh" is frequently spoken in everyday conversation. People often respond in various situations by saying only "eh" and listeners understand. In a previous study the F0 contours of emotional speech as spoken on the interjection of "eh" were described. The results showed that duration and slope, as calculated from the range of F0 divided by duration, along with F0 rising and falling characteristics, tend to vary significantly for each emotion [6].

In the study, stimuli were categorized in terms of their F0 characteristics: duration, slope, and rising/falling F0. These parameters can describe the four emotions of "wonder," "disappointment," "asking again," and "affirmation." However, for "doubt," "postponement" and "hesitation," these three parameters are not sufficient.

A subsequent study investigated whether listeners could identify the intended emotion of speakers' productions of "eh" presented out-of-context of the original dialogue [7]. All the emotions except "hesitation" had a high percent of correct answers. Responses, which were not consistent with the original context, can be accounted for by their F0 contour characteristics, especially by duration and slope. These results suggested that F0 contours have sufficient information for listeners to be able to recognize emotions without dialogical context.

From these two experiments, it seems that duration and the slope of F0 contour is related with emotion recognition. However, the stimuli used in the experiments were naturally pronounced voices without any processing. Thus, other parameters such as intensity or voice quality were included in the stimuli. To ensure the effect of duration and slope of F0, we used the stimuli which had differences only in duration and F0, through editing on computer software. We modified the duration and F0 gradually to investigate its effect on estimating a speaker's emotional state.

In addition to editing stimuli, we improved the procedure of the experiments in the recognition test. The interjection "eh" was used in various situations, such as in cases of "surprise," "disappointment," "affirmation," or "hesitation." The previous study [6,7] proposed seven contexts, "asking again," "affirmation," "postponement," "wonder," "doubt," "disappointment," and "hesitation." However, these 7 contexts were not complementary contexts. The following dialogue, used in the experiments, illustrates some of these contexts.

A: Oh, your name isn't on that list.
You failed the entrance examination.
B: ***Eh***.

The "eh" in this dialogue is supposed to be in the context of "disappointment," but it also implies

"affirmation." It cannot be helped that the listeners thought the stimuli were "affirmation."

In this experiment, we made the contexts complementary to each other, (1) disagreement (surprise, wonder, doubt), (2) agreement (affirmation, nodding), and (3) hesitation (postponement). We asked the listeners to judge if the stimuli were suited to the context or not in order to make the response clear.

2. Method

2.1. Participants

Ten Japanese adults (7 Male), with a mean age of 35, participated in the experiment. All participants had normal hearing.

2.2. Stimuli

The stimulus material was one syllable, [e], pronounced by Japanese a male. He was instructed to say the sound, which means "shaft" or "helve" in Japanese, without any special emotions. The original sound was 189ms long; with a mean F0 of 131Hz (Figure1). The sound was a single vowel syllable, so that the spectral information was constant, and the same waveform was repeated. First, we selected one cycle of the waveform, where the power was at its maximum within the utterance. We then copied that part into the same point to make longer syllables. We made four types of different durations, 189ms (original), 418ms, 639ms, and 868ms syllables. By a preliminary experiment, we determined that 868ms is the maximum length in which this sound can be perceived as the word "eh." We used audio editing software named Cool Edit 2000 for this editing.

Next, we modified the frequency. All the stimuli had 131Hz F0 as a mean. They were flat sounds without any pitch changes (Figure2). Then, we modified the flat sounds. Pitch ranges were, 25Hz, 50Hz, 75Hz, 100Hz, and 125Hz. All of them had two patterns: a rising tone and a falling tone. Changes were linear (Figure 3). We made 44 stimuli (11 slope x 4 duration) all together.

The stimuli (wave format) were played on a personal computer (DELL INSPIRON 8100), and were presented through a loudspeaker (SONY SMS-1P) placed in front of the participants. The audio was set at a comfortable listening level (about 60dB-A, peak with fast scale from a sound level meter). The floor noise level was about 30dB(slow, B).

2.3. Procedure

Each stimulus was presented once every five seconds in a random order. It took about 10 minuets for each participant. The participants were told that the voices were a man's voice (not a woman), who is most likely talking on the phone. The participants were asked to detect "yes" or "no" to each emotional state presented beforehand, after listening to the sound. Before the presentation of the stimuli, one of the three kinds of emotional states, "disagreement," "agreement," and "hesitation" was presented visually on the display.

We explained beforehand that by "disagreement" we meant one of three things: (1) surprise, (2) someone who cannot believe what another has just said, or (3) the results were incongruent with what the listener had predicted. Two meanings of "agreement" were (1) affirmation, or (2) back-channel feedback (such nod). Finally, three meanings of "hesitation" were (1) confusion, (2) cannot decide immediately, and (3) looking for a word.

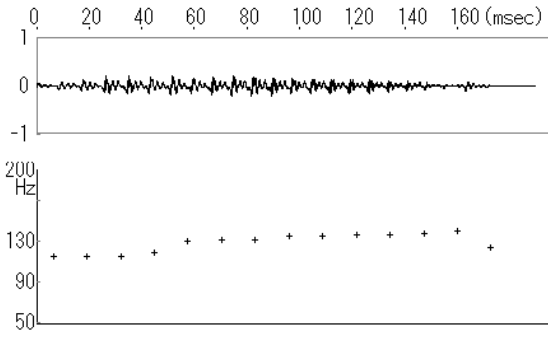

Figure 1*: Waveform and F0 contour of original voice.*

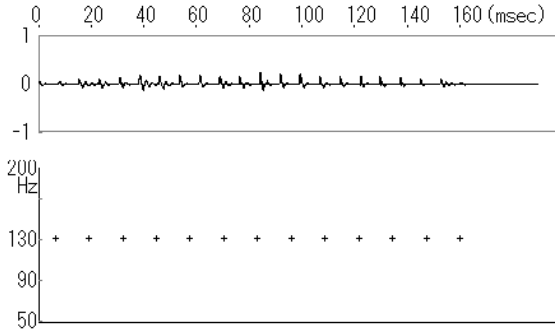

Figure 2: *Waveform and F0 contour after modifying the sound into F0 range 0 (flat).*

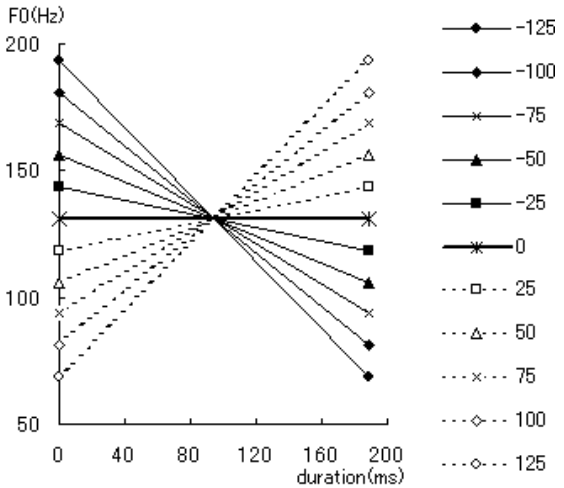

Figure 3*: Eleven different F0 contours (duration: 189ms). The numeral shows the range of F0 (ending F0 minus beginning F0).*

Before the experiment, we did a practice session. We used the natural speech of a man, who was the speaker in previous studies [6,7]. The purpose of the practice was to get the listener acquainted with the comprised answering system. All participants finished the practice session without any difficulties. The experiment followed, where every participant had 131 randomly ordered (11 slope x 4 duration x 3 contexts)

3. Results

We divided the results into three emotional states, and calculated the recognition rate. There is statistical difference between rising tones and falling tones ($F(10,90)=5.723$, $p<.01$). Figure 4 to 6 show these differences. The horizontal axis shows the range of F0. The value was calculated as ending F0 minus beginning F0. The vertical axis is the percentage of recognition as the emotion. There are four lines indicating the duration.

3.1. Recognition of "Disagreement"

Figure 4 shows the recognition rate for the question "Disagreement." It is important if the sounds are rising tone (F0 range = plus) or falling tones (F0 range = minus), but not so important if there are different values in the duration and F0 ranges.

85.0% of the rising tone stimuli were recognized as "disagreement." The percentages of recognition were above 70% for all stimuli, regardless of their duration. 70% were far above the 50% chance level. On the other hand, the average of the recognition rate for falling tone stimuli were only 15.5%. It seems important if the sound is rising or falling, in the case of recognizing "disagreement." There is a significant difference between rising tones and falling tones ($F(1,9)=70.43$, $p<.01$).

We then analyzed the data by its duration. 52.0% of responses for the stimuli with 189ms duration were recognized as "disagreement." In the same way, 52.0% with 418ms, 49.0% with 639ms, and 48.0% with 868ms. There was no significant difference when looking at variations in duration ($F(3,27)=0.383$, n.s.).

Next, we analyzed F0 range differences. For rising tone stimuli, 82.5% were recognized as "disagreement" with the range of 25Hz, 87.5% with 50Hz, 80.0% with 75Hz, 85.0% with 100Hz, 90.0% with 125Hz. There was no significant difference among F0 ranges ($F(4,36)=0.474$, n.s.). For falling tone stimuli, 22.5% were recognized as "disagreement" with the range of -25Hz, 20.0% with -50Hz, 7.50% with -75Hz and -100Hz, and 20.0% with -125Hz. There was no significant difference among the F0 ranges, either.

3.2. Recognition of "Hesitation"

Figure 5 shows the recognition rate for the question "hesitation. The response for rising tone stimuli and falling tone stimuli were significantly different ($F(1,9)=10.618$, $p<.01$).

For rising tone stimuli, when the range of F0 was bigger, the recognition rate went lower ($F(4,36) = 6.908$, $p<.01$). This merits closer examination. An F0 range of 25Hz was recognized as "hesitation" more than 75Hz, 100Hz, 125Hz, but no different from 50Hz. An F0 range of 50Hz was higher than 100Hz, 125Hz, but no different from 75Hz. There were no significant differences among 25Hz, 50Hz, and 75Hz. The important result here shows, most of the stimuli were below a 50% chance level, but 868ms with –25Hz, 639ms with –25Hz, 639ms with –50Hz. That is, when the sound is a rising tone, people will not recognize it as "hesitation", even when its duration is long.

For falling tone stimuli, the recognition rate, as "hesitation" became higher when the duration was longer ($F(3,27) = 15.779$, $p< .01$). The average rate for each duration was 16% with 189ms, 32% with 418ms, 68% with 639ms, and 78% with 868ms. On the other hand, the F0 range is not important for recognizing "hesitation."($F(4,36) = 0.953$, n.s.)

3.3. Recognition of "Agreement"

Figure 6 shows the recognition rate for the question "Agreement." All the rising tone stimuli had responses under 50% chance levels. It seems that the sounds with rising tone, will not to be recognized as "agreement" regardless of their duration or F0 range.

In contrast, the recognition rate goes higher when the F0 range becomes bigger in the case of falling tone stimuli ($F(4,36)= 5.925$, $p<.01$). It is important if the sounds are rising or falling tones, but not so important if duration and F0 ranges vary.

85.0% of the rising tone stimuli were recognized as "disagreement." The percentages of recognition were above 70% for all stimuli, regardless of their duration. 70% is far above the chance level. On the other hand, the average of the recognition rates for falling tone stimuli were only 15.5%. It seems important if the sound is rising or falling, in the case of recognizing "disagreement." It is significantly different between rising tones and falling tones ($F(1,9)=70.43$, $p<.01$).

Analyzed by duration, the 52.0% of responses for the stimuli with 189ms duration were recognized as "disagreement. In the same way, 52.0% with 418ms, 49.0% with 639ms, and 48.0% with 868ms. There was no significant difference about duration ($F(3,27)=0.383$, n.s.).

Analyzed by F0 range differences, rising tone stimuli, 82.5% was recognized as "disagreement" with the range of 25Hz, 87.5% with 50Hz, 80.0% with 75Hz, 85.0% with 100Hz, 90.0% with 125Hz. There was no significant difference among F0 ranges ($F(4,36)=0.474$, n.s.). For falling tone stimuli, 22.5% were recognized as "disagreement" with the range of -25Hz, 20.0% with -50Hz, 7.50% with -75Hz and -100Hz, and 20.0% with -125Hz. There was no significant difference among the F0 ranges, either.

4. Discussion

Sounds with a rising tone are recognized as "disagreement," regardless of the range of F0. Further, these sounds would never be recognized as "agreement." The sounds with a small F0 range (25Hz or 50Hz) and with a long duration are often recognized as "hesitation." Other sounds (i.e., big F0 range or short duration) are never recognized as "hesitation."

The stimuli with a falling tone are recognized as (1) "hesitation", when the sounds have a long duration, and (2) "agreement," when the sounds have a short duration. The

falling tone stimuli are never recognized as "disagreement."

The sounds with falling tones are more recognized as "agreement" when the F0 ranges get larger. Generally, recognition for "hesitation" gets higher when duration gets longer. However, there is a possibility that even if the F0 range is big enough, some sounds will not recognized as "hesitation." The reason is, when the F0 range is -125Hz, the recognition rate for "hesitation" goes down than F0 range -100Hz. We guess that the recognition rate will get lower when the F0 ranges are too big. To investigate this hypothesis, we will have to do another experiment with a bigger F0 range than we used here.

The purpose of this study is to ensure the effect of the "duration" and the effect of the "slope of F0 contour," which were focused on in the previous studies [6,7]. In what context (in the speakers' emotional state) do these factors become effective? Are these factors effective even if the sound has no other information like voice quality? We changed the F0 range as a parameter to investigate the slope of the F0 contour in this experiment. We could confirm that people can sufficiently distinguish "disagreement," "hesitation," or "agreement" only from duration and F0 ranges.

From the results of this study, we can guess that people may be using prosodic information, not segmental information. In that case, we must be able to obtain similar results when using sounds that have not only prosodic information but also segmental information. For that purpose, we are now preparing further experiments, using triangle waves, as stimuli.

5. Acknowledgements

This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grant-in-Aid for Young Scientists (Category B), 15700194, 2003-2005.

6. References

[1] Fujisaki, H., 1997. Prosody, models, and spontaneous speech. In Sagisaka, Y. et al. (editors), *Computing Prosody: Computational Models for processing Spontaneous Speech*, Springer, 27-42.

[2] Scherer, K. R., 1986. Vocal affect expression: a review and a model for future research. *Psychological Bulletin*, 99, 143-165.

[3] Utsuki, N., 1979. Recognition of vocally expressed emotion. In *Proceedings of the 43rd Conference of Japanese Psychological Association*, 387.

[4] Maekawa, K., 1998. Phonetic and phonological characteristics of paralinguistic information in spoken Japanese. In *Proceedings of ICSLP 1998 The 5th International Conference on Spoken Language Processing*, 3, 635-638.

[5] Eda, S., 2000. Identification and discrimination of syntactically and pragmatically contrasting intonation patterns by native and non-native speakers of standard Japanese. In *Proceedings of ICSLP 2000 The 6th International Conference on Spoken Language Processing*, 2, 361-364.

[6] Hayashi, Y., 1998. F0 contour and recognition of vocal expression of feelings: using the interjectory word "eh". *Technical report of IEICE SP*, 43, 65-72.

[7] Hayashi, Y., 1999. Recognition of vocal expression of emotions in Japanese: using the interjection "eh". In *Proceedings of ICPhS'99, the 14th International Congress of Phonetic Sciences*. San Francisco, 3, 2355-2358.

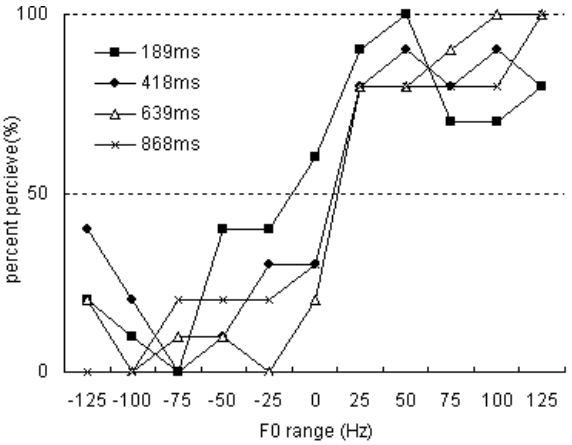

Figure 4: Recognition rate for "disagreement."

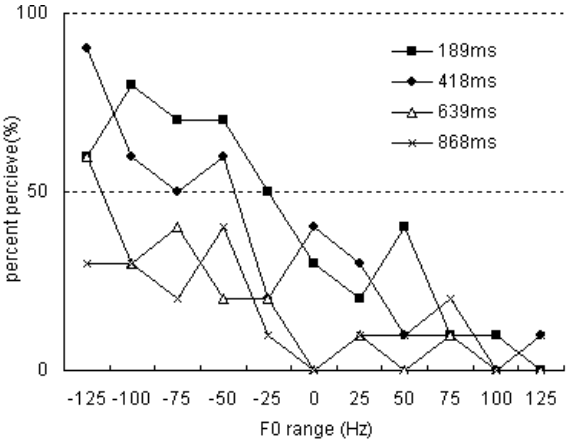

Figure 5: *Recognition rate for "hesitation."*

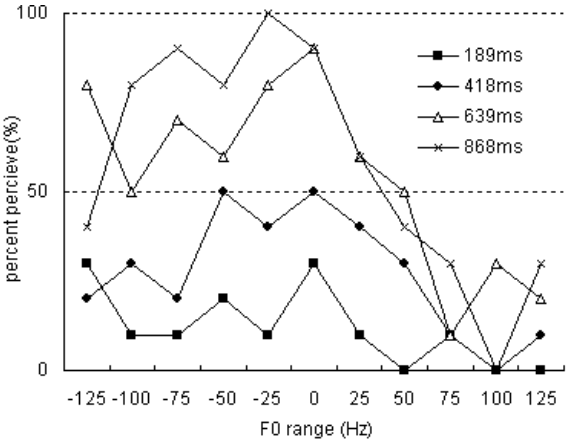

Figure 6*: Recognition rate for "agreement."*