Detection of Baby Voice and its Application Using Speech Recognition System and Fundamental Frequency Analysis

SHOTA YAMAMOTO[1], YASUNARI YOSHITOMI[2], MASAYOSHI TABUSE[2], KOU KUSHIDA[3], AND TARO ASADA[2]
1: Kumon Educational Japan Co., Ltd.
5-6-6 Nishinakashima Yodogawa-ku Osaka 532-8511 JAPAN
2: Graduate School of Life and Environmental Sciences, Kyoto Prefectural University
1-5 Nakaragi-cho Shimogamo Sakyo-ku Kyoto 606-8522 JAPAN
yoshitomi@kpu.ac.jp, http://seika.kpu.ac.jp/~yoshitomi/
3: Kyoto Prefectural Tanabe Senior High School
24 Kawara Kamitani Kyotanabe Kyoto 610-0361 JAPAN

Abstract: - We propose a method for detecting a baby voice using a speech recognition system and fundamental frequency analysis. We propose the following two conditions for recognizing a sound form segment of a baby voice. Condition 1: The word reliability for a sound form segment obtained by using Julius is under a threshold, Condition 2: For a certain time period, the fundamental frequency of the sound form segment changes by another threshold or over. When at least one of the above two conditions is met, the sound form segment is judged as coming from a baby voice. We successfully applied the proposed method to pattern recognition of a baby's emotion.

Key-Words: - Baby Voice, Speech Recognition System, Fundamental Frequency, Emotion, Pattern Recognition, Baby Care Support

1 Introduction

In Japan, the number of births has shown a downward trend. This trend is considered a serious problem. One of the reasons for the decrease in births may be the lack of child care support as a social system.

The number of persons living together has also shown a downward trend. As a result, a young mother often takes care of her baby by herself. She might not have the opportunity to learn appropriate child care. In some cases, the mother works outside the home and needs a daycare to take care of her baby.

As is well known, it can be extremely dangerous to leave a baby alone. However, the mother typically must do a lot of housework and may want to enjoy another activity or hobby. In this case, raising a child safely becomes more difficult. Unfortunately, baby abuse and neglect have been increasing because of the stress of a mother taking care of her baby. Therefore, we think that decreasing the mother's stress may slow the decrease in the number of births.

Investigations on a baby voice mainly using frequency analysis have been reported [1]–[9]. However, form clipping of the sound form of a baby voice from wave data has been performed by manual operation. Because manual operation is time-consuming, automatic recognition of a baby's emotion, which needs automatic detection of a baby voice, is still not resolved.

We have researched a system for improving baby care by recognizing a baby's emotion from the baby voice. As stated above, form clipping of the sound form of a baby voice from wave data has typically been performed manually [9]. In this study, we propose a method for detecting a baby voice using a speech recognition system [10] and fundamental frequency analysis. We successfully applied the proposed method to pattern recognition of a baby's emotion.

2 Proposed Method

Fig. 1 shows the flow chart of our method. We explain the procedure in the following.

2.1 Word Recognition

We use a speech recognition system named Julius [10] for word recognition. When Julius recognizes a sound form segment as expressing a certain word, the sound form segment is used in the next processing. However, in the proposed method, when Julius recognizes a sound form segment as silent, the sound form segment is omitted.

2.2 Threshold Treatment of the Sound Form

The sound form segment recognized as not silent by Julius may contain sound mostly composed of noise.

We neglect the sound form segment having amplitude below a threshold, as decided experimentally beforehand, because the sound form segment may be composed of noise. The threshold was experimentally decided as the value of $m + 2\sigma$, where m and σ are the average and standard deviation, respectively, of the sound amplitude when neither voice or unusual sound is present in the recorded sound form.

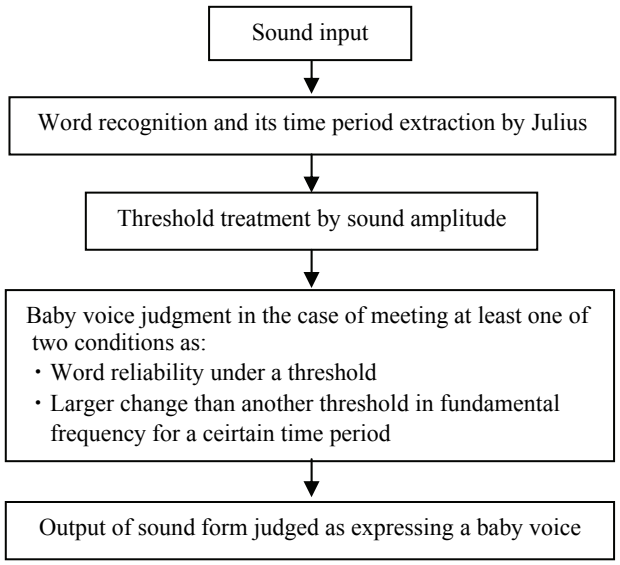

Fig. 1. Flow chart for detecting sound form describing a baby voice.

The sound form segment extracted by the method described in Section 2.1 and simultaneously having amplitude over the threshold is a candidate of being a baby voice.

2.3 Detection of Baby Voice

We propose the following two conditions for recognizing a sound form segment as coming from a baby voice:

Condition 1:
The word reliability for a sound form segment obtained using Julius is under a threshold.

Condition 2:
For a certain time period, the fundamental frequency of the sound form segment changes by another threshold or over.

The two thresholds and the time period, which are described above, are experimentally decided. When at least one of the above two conditions is met, the sound form segment is judged as coming from a baby voice.

2.3.1 Judgment by Word Reliability

Julius outputs word reliability of 0 to 1 for a recognized word. High word reliability means that the difference of likelihood to other words as candidates is large. Accordingly, when the value of the word reliability is small, the speech recognition result may be doubtful. In the case of a baby voice, the word reliability given by Julius is assumed to be small.

We decided a threshold for distinguishing the word reliability of a baby voice from that of an adult.

2.3.2 Fundamental Frequency Analysis for a Short Time

In this study, we found that the fundamental frequency of a baby voice suddenly changes, for example, in the time ranges of 1.5 to 1.8 s and 6.6 to 6.8 s, as shown in Fig. 2.

Therefore, we decided a certain time period and another threshold of change in the fundamental frequency for distinguishing a baby voice from other sounds.

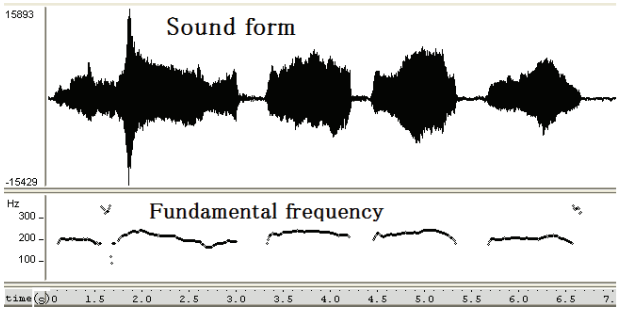

Fig. 2. Large change of fundamental frequency of baby voice for a short period.

3 Performance Evaluation

3.1 Experimental Conditions

3.1.1 Voice Recording

We used a wireless or wired microphone and notebook computer for recording sound. The sound form is saved as WAVE files with the following specifications: PCM format, 16 kHz, 16 bits, monaural.

We used three kinds of sound forms for a total of four experiments: (1) voice of male baby one and one-half months old, which was recorded in his home and composed of 57 pieces of continuous sound forms having 10 to 12 cries for each, (2)-1 a mixture (hereinafter referred to as mixture 1) of baby and adult voices, which were recorded in a day nursery and composed of three pieces of continuous sound forms

having 8 to 21 voice regions for each voice, (2)-2 a mixture (hereinafter referred to as mixture 2) of baby and adult voices, which were also recorded in the day nursery and composed of five pieces of continuous sound forms having 8 to 21 voice regions for each, (3) voices of five male and two female adults, which were recorded in our laboratory at Kyoto Prefectural University and composed of seven pieces of continuous sound forms having 14 to 22 pronunciations of "taro" for each.

The sound data described as (3) was the same as those of the neutral part in our previously reported study [11].

The experiment was performed in the following computational environment: the personal computer was a DELL OPTIPLEX GX260 (CPU: Pentium IV 2.4 GHz; main memory: 512 MB); the OS was Microsoft Windows XP; the development language was Microsoft Visual C++ 6.0.

3.1.2 Method for Evaluation

First, we obtained the threshold for the amplitude of the sound form and the threshold for word reliability using sound data for the first half of (1) (baby voice) and (3) (adult voices) described in Section 3.1.1. Fig. 3 shows the distribution of the word reliability of baby and adult voices. Then, using the sound data of (1) (baby voice), we experimentally decided the condition of sudden change in the fundamental frequency as 0.1 s of time for checking and 150 Hz as the threshold.

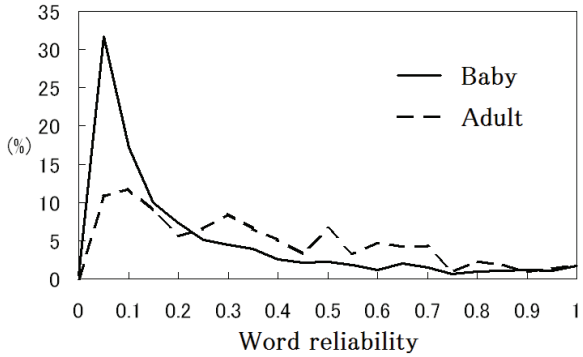

Fig. 3. Word reliability distributions of baby and adults.

Moreover, to judge whether the sound form segment came from the baby voice, we checked all sound form segments, which were recognized as not silent by Julius, in the following two ways: the sound form segments were checked by eye, and its real sound was checked by ear. Then, we compared the sound form segments judged as coming from the baby voice by checking by eye and ear to those by the proposed method. Here, we used two criteria, detection and mis-detection rates defined below, for evaluation of the proposed method.

The detection rate (%) is defined as $(N(A \cap B) / N(A)) \times 100$, where N(X) is the number of elements of set X, A is a set of sound segments of the baby voice, and B is a set of sound segments detected by the proposed method. The mis-detection rate (%) is also defined as $\{(N(B) - N(A \cap B)) / (N(J) - N(A))\} \times 100$, where N(X), A and B are described above, and J is the set of all sound form segments judged as being not silent by Julius. Fig. 4 shows the relationship among the sets of A, B, and J.

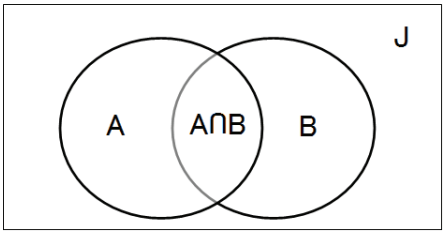

Fig. 4. Schematic diagram showing sets and their relationship used for explaining detection and mis-detection rates.

Next, we used the proposed method for emotion recognition of the voice of the male baby one and one-half months old. Then, the result was compared with those produced by manual sound form clipping. The pattern of emotions assumed was "discomfortable", "hungry", and "sleepy". The emotion expressed by his cry was recognized by his caregiver in recording his cry. The first half of the recorded sound form was used as the training data, and the second half was used as test data for evaluating the application potential of the proposed method.

Fig. 5 shows a flow chart for emotion recognition of a baby using his or her voice. The 32-dimensional Fast Fourier Transform (FFT) was performed for the sound form clips as training data. Then the power of the sound form judged as a silent region was subtracted from each power of frequency element. The power of each frequency element after the subtraction was treated as one of the elements of the feature vector. We performed Principal Component Analysis (PCA) for the feature vectors of training data under the condition that the main components were assembled up to the lowest dimension, at which the accumulated contribution ratio exceeded 80%. Then, the feature vectors for the sound form clips of the test data were calculated in the same way as those for the training data. The emotion of the baby was recognized by the nearest neighbor

criterion applied to the feature vector obtained from the test data of the sound form clip after projecting the feature vector on the PCA space set from the training data, as described above. Then the emotion with the highest frequency among the recognition results for the sound form clips was judged as the emotion expressed by the baby's cry. When the recognition results of two or more emotions had the highest frequency among those of the three emotions, the second nearest neighbor feature vector for each sound form clip in the PCA space was treated in the same manner as that of the nearest one for judging the emotion expressed by the baby's cry.

Table 1 shows an example that explains the judgment of the emotion expressed by the baby's cry. In this case, 'hungry' is judged as the emotion expressed by the baby's cry because the emotion with the highest frequency is 'hungry' among the three emotions.

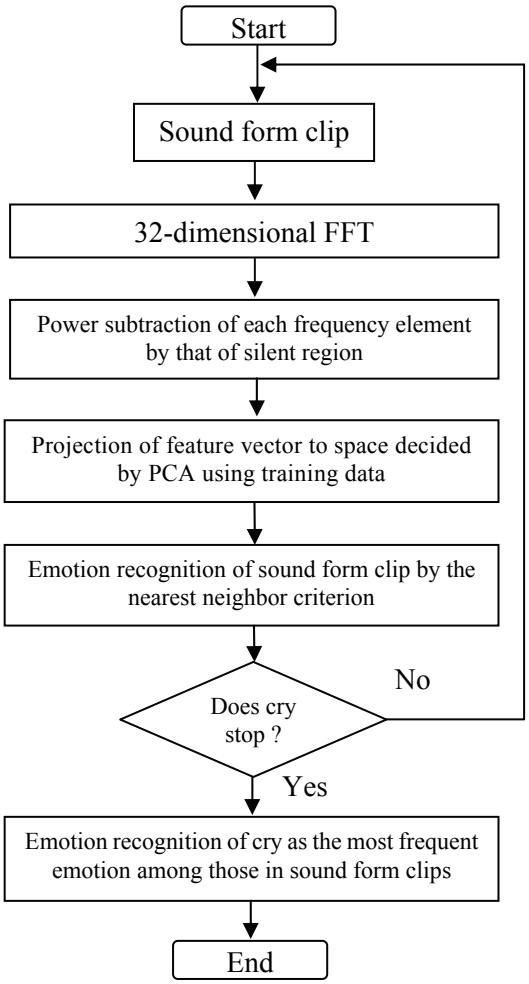

Fig. 5. Flow chart for recognizing baby's emotion from his or her voice.

Table 1. Example of emotion recognition using continuous voice form.

Voice form clip No.	Discomfortable	Hungry	Sleepy	Judgment
1			○	Hungry
2	○			
3		○		
4			○	
5		○		
6	○			
7		○		
8		○		

3.2 Experimental Results and Discussion

3.2.1 Baby Voice Detection

Table 2 shows the results of the baby voice detection. The detection rates fall in the relatively narrow range of 65.0% to 69.4%, whereas mis-detection rates are spread in a range of 2.0% to 34.1%. In the case of mixture 1, the mis-detection rate is high because it contains much noise, such as that of an automobile and a toy for calming the baby. Noise causes low word reliability by Julius, resulting in a high mis-detection rate. Moreover, in both cases of mixtures 1 and 2, the adult voice for calming the baby causes low word reliability, resulting again in a high mis-detection rate. As shown in Fig. 3, word reliability for the adult voice of "taro" is sometimes low. Therefore, the adult voice may lead to mis-detection because word reliability is one of the judgment criteria for a baby voice.

The main target of this study to apply the proposed method is to improve baby care support in the baby's home. Therefore, the most important result among those shown in Table 2 is the column of "Baby". In this case, the detection rate is 69.4% and the mis-detection rate is 2.0%. Accordingly, the proposed method can successfully support baby care.

Table 2. Results of baby voice detection.

	Baby	Adult	Mixture 1	Mixture 2
Number of sound clips detected by Julius	1172	333	186	239
Number of baby voice clips	569	0	60	126
Number of sound clips detected by proposed method (Number of baby voice clips in the above number)	407 (395)	66 (0)	82 (39)	97 (94)
Detection rate	69.4%	——	65.0%	66.2%
Mis-detection rate	2.0%	19.8%	34.1%	2.7%

We think that the stress of baby care while doing housework away from the baby and waiting for

notification of the baby's cry by using the proposed method would be less than the stress while doing housework and always paying attention to sound sent through a wireless microphone beside the baby. The system can have the function sending an image of the baby to the person taking care of the baby only when the baby is crying. Consequently, the system enables the person to stay apart from the baby when the baby is not in need of immediate care.

3.2.2 Application to Emotion Recognition

Table 3 shows the emotion recognition results using sound form clips extracted by the proposed method or by manual operation. The average accuracies of emotion recognition are 62.1% and 66.6% for the sound form clips extracted by the proposed method and by manual operation, respectively. This result suggests that the proposed method is applicable to emotion recognition as a preprocessing module with almost the same ability as that of manual clipping of the sound form.

The accuracy of "discomfortable" is lower than that of other types of emotion. This is because the sound expressing "discomfortable" has significant variations, resulting in difficulties of pattern recognition when the nearest neighbor criterion in the PCA space is used.

Table 3. Results of emotion recognition rate using continuous voice form.

Baby voice form clip / Emotion	Proposed method	Manual operation
Discomfortable	3/10(30%)	5/10(50%)
Hungry	13/14(92.9%)	10/15(66%)
Sleepy	2/5(40%)	5/5(100%)
Total	18/29(62.1%)	20/30(66.6%)

One of the reasons why the emotion recognition rate of the proposed method for "sleepy" is lower than that by manual operation is the low detection rate for the sound form clip expressing "sleepy". This is because the sound form clip expressing "sleepy" may not be detected by the proposed method, which uses a threshold for the amplitude of the sound form to neglect noise. Because the sound form clip expressing "sleepy" tends to have small amplitude, this emotion is sometimes neglected by the proposed method.

4 Conclusion

We propose a method for detecting a baby voice using a speech recognition system and fundamental frequency analysis. The main target of this study to apply the proposed method is to improve baby care support in the baby's home. In the case, the detection rate of baby voice is 69.4% and the mis-detection rate is 2.0%. We successfully applied the proposed method for pattern recognition of a baby's emotions.

References:

[1] H. M. Truby and J. Lind, Cry Sounds of the Newborn Infant, *Acta Paediatr. Scand. Suppl.*, Vol. 163, 1965, pp. 8-59.

[2] P. H. Wolff, The Natural Histry of Crying and Other Vocalizations in Early Infancy, in B. W. Foss, (ed.), *Determinants of Infant Behavior IV*. London: Methuen and Co, 1969, pp.81-111.

[3] B. F. Fuller and Y. Horii, Differences in Fundamental Frequency, Jitter, and Shimmer among Four Types of Infant Vocalizations, *J. of Communication Disorders*, Vol. 19, No. 6, 1986, pp. 441-447.

[4] B. F. Fuller and Y. Horii, Spectral Energy Distribution in Four Types of Infant Vocalizations, *J. of Communication Disorders*, Vol. 21, No. 3, 1988, pp. 251-261.

[5] Q. Xie, R. K. Ward, and C. A. Laszlo, Automatic Assessment of Infants' Levels-of-distress from the Cry Signals, *IEEE Trans. Speech and Audio Processing*, Vol. 4, No. 4, 1996, pp. 253-265.

[6] K. Kikuchi and K. Arakawa, Estimation of the Cause of Younger Babies' Cries by Frequency Analyses of Their Voice, *IEICE Technical Report*, Vol. SIS2005-69, 2006, pp. 55-60 (in Japanese).

[7] Y. Mima and K. Arakawa, Cause Estimation of Younger Babies' Cries from the Frequency Analyses of the Voice – Classification of Hunger, Sleepness, and Discomfort, *IEICE Technical Report*, Vol. SIP2006-79, 2006, pp. 43-46 (in Japanese).

[8] K. Nishimura, Y. Mima, and K. Arakawa, Estimation of the Cause of Cries for One-Month Old Babies, *IEICE Technical Report*, Vol. SIS2006-87, 2007, pp. 35-40 (in Japanese).

[9] K. Kushida, M. Tabuse, and Y. Yoshitomi, Pattern Recognition of Emotional States in Baby Voice, *Proc. of Human Interface Sympsium 2008*, pp.25-28, 2008 (in Japanese).

[10] http://julius.sourceforge.jp/

[11] M. Nakano, F. Ikezoe, M. Tabuse, and Y. Yoshitomi, A Study on the Efficient Facial Expression Using Thermal Face Image in Speaking and the Influence of Individual Variations on Its Performance, *J. IEEJ*, Vol. 38, No.2, 2009, pp. 156-163 (in Japanese).