COLLECTION AND ANALYSIS OF EMOTIONAL SPEECH FOCUSED ON THE PSYCHOLOGICAL AND ACOUSTICAL DIVERSITY

Takahiro Miyajima[a], *Hideaki Kikuchi*[a] *& Katsuhiko Shirai*[b]

[a]Faculty of Human Sciences, Waseda University, Saitama, Japan;
[b]Faculty of Science and Engineering, Waseda University, Tokyo, Japan
miyajima@toki.waseda.jp

ABSTRACT

How to effectively collect diverse data of spontaneous speech and to upgrade various speech processing techniques is a serious issue. In this paper, we introduce our proposed method "SEN method", which aims to collect psychologically and acoustically diverse acted speech effectively with naturalness. We created detailed directions, on the basis of various real-life scenarios, consulting with a professional actress in order to facilitate duplicating diverse expressions for the actor or actress. We compared fifty speech data by the SEN method and fifty others by the legacy method, where simple basic emotional words are used as prompts. In the psychological space, the SEN data filled up low density areas of the space of the legacy method. In order to confirm the causes of this phenomenon, we analyzed the relationship between the psychological and acoustical features. Our results demonstrate the advantage of the SEN method, which is the generation of psychologically diverse speech that cannot be described by representative acoustical features.

Keywords: emotional speech, diversity, acting script

1. INTRODUCTION

Techniques for processing information regarding human emotions have recently been improved. Erickson [2] reviewed many studies on emotional speech, and the problems that are encountered from the viewpoint of "expressive speech" and indicated four problems that constitute future research challenges: (1) data collection, (2) data labeling, (3) analysis of voice quality, and (4) expressive speech synthesis.

Our first concern is the construction of a general purpose corpus for emotion research, as a solution to the data collection issue. Ideally, the essential conditions for a general-purpose corpus are that the data must be gathered from many people in many contexts, it must be spontaneous, and it must express diverse emotions. Some recent studies such those by as Campbell [1] and Steidl [5] have suggested a new approach to collect unintentional and spontaneous speech: All expressions of speech generated in daily life or in imitations of daily life over a period of several days to several weeks are recorded. However, there remain some problems with limited diversity in the method of collecting spontaneous speech from daily life. It is easy to imagine that the data will be biased, psychologically and acoustically, without any control, even if significant time is spent in collection. Fully understanding the importance of recording spontaneous speech from daily life, we suggest a new method that uses professional voice actors and an "acting script." Thus, the aim of our study is to effectively collect psychologically and acoustically diverse speech in support of our final goal, which is to construct a general-purpose corpus for emotion research.

2. METHODLOGY

2.1. Definitions of impression and emotion

Moriyama [3] indicated that speech catalyzes both a subjective emotion (mental state) and an objective emotion (assumption of the receiver of the impulse), and he settled upon the latter as the goal of his research on emotions. The principles of our study are based upon Moriyama's definition. We also define the improvement of the psychological diversity of an emotional speech database as being equal to the expansion of the psychological space for emotions. There are two possible expansions to this: (a) spread of spatial dimensions, (b) populating low-density areas of existing space.

2.2. Procedure of data collection

In this chapter, we describe the sequence involved in our method (Figure 1).

First, we considered what type of information is required for an actor to produce diverse expressions of speech. We refer to such information as

"Format of Acting Script." We discussed various items (such as the personality, the context, and so on) and selected a number of significant items that would concern a professional voice actor. Table 1 presents the items we adopted and examples of each item. We denominate to the speech data, collected by our proposed method, as "SEN (meaning "Thousand" in Japanese) Speech Data (=SEN data)." For concrete content of script, we selected various expressions from TV dialogues, as these programs contained diverse speech. To make our selections, we recorded TV programs from one Japanese broadcasting station for 24 h. We selected 304 expressions and produced the same number of acting scripts from 24 h of recorded data, aiming to generate a complete diversity of speech for our scripts.

Figure 1: Concept of proposed method.

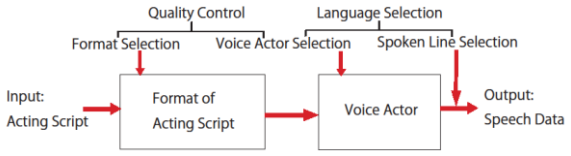


Table 1: Format and example of acting script.

Format		Example
Common	Location/situation	Drama, hall in dormitory
	Relationship	Same school and dormitory, good friends
Listener	Age/gender	Seventeen or eighteen, male
	Career	High school student
	Portrait	Noisy, cluttered
Speaker	Age/gender	Seventeen, female
	Career	High school student
	Portrait	Pretending to be man, entering boy's school, ataraxia, insensitive to love
	Context/Background	Amazed at surrounding commotion

We then decided upon the actual lines to be spoken. We selected "/aH,so'Hdesuka/" (meaning "Oh, I see" in English) as the spoken lines. The reasons for this are as follows: (1) an actor could convey a variety of paralinguistic information by changing the prosody, because the linguistic information in the sentence has multiple meanings, (2) the meaning of the passage is neutral and will not influence the evaluation of the impression, (3) it includes an accent phase that enables the expression to be varied more easily, (4) it has a high frequency of appearance in daily conversation.

We recruited a female professional voice actress whose career included eighteen years of voice acting and theatrical performance. We extracted one hundred acting scripts in which the gender of the speaker is female and recorded speech data.

2.3. Hypothesis and proof

Our hypothesis is that psychological/acoustical features become diverse as prompts given to actor/actress become diverse and detailed. To collect evidences of this hypothesis, it is absolutely necessary to consider what the baseline of diversity of human speech is. There, we prepared speeches collected by very simple method using "basic emotion words" as data for comparison. And we define such speech data "Typical data".

We implemented the recording of Typical data using eight sets of basic emotion words (table 2), referring to Ortony & Turner [4]. The total number of basic emotions was forty-nine, and we added an additional word, "athymia," to the word set. The speech actor used was the same person used to obtain the SEN data.

Table 2: Summary of basic emotions by Ortony & Turner [4].

Proposer/Year	Basic Emotional Words
Arnold(1960)	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness
Ekman, et al.(1982)	Anger, disgust, fear, joy, sadness, surprise
Frijda(1968)	Desire, happiness, interest, surprise, wonder, sorrow
James(1884)	Fear, grief, love, rage
McDougall(1926)	Anger, disgust, elation, fear, subjection, tender-emotion, wonder
Mowrer(1960)	Pain, pleasure
Oatlay, et al.(1987)	Anger, disgust, anxiety, happiness, sadness
Plutchik(1980)	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise

3. ANALYSIS OF PSYCHOLOGICAL FEATURES

In this chapter, we compare the psychological features of the SEN data and the Typical data, to confirm that the proposed method advances the diversity of emotional speech corpus.

We adapted Moriyama's "emotion express words [3]" for the evaluation. They indicated that highbred assessment words which are needed for effective evaluation of objective emotions in human speech. They picked nine emotional words (see "word" column in table 4) as factors, drawn from forty-six emotional words in various existing works. We implemented the impression evaluation using these emotional words, as well as the naturalness of speeches questionnaire.

Six male and six female university students served as the evaluators. We picked fifty speech data samples from the SEN data, in order to ensure a balance with the number of Typical data samples. Evaluators listened, at random, to the one hundred speech data samples as stimuli, and rated each on a seven-level evaluation scale for emotional words.

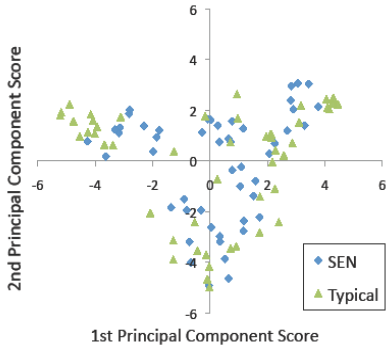
The evaluation scale was as follows: 1 implied not included at all, 4 was neutral, and 7 indicated included very much.

The "Mean" and "Stddev" columns in Table 3 show the means and standard deviations of the evaluation scores of the one hundred Typical and SEN data samples (fifty each), by each emotional word.

Table 3: Evaluation parameters of psychological analysis.

Emotional Words	Mean		Stddev		Eigenvectors		
	Typical	SEN	Typical	SEN	PC1	PC2	PC3
Anger	2.93	2.71	1.84	1.33	-0.57	0.26	-0.23
Pleasure	2.53	2.58	1.53	1.33	0.40	0.35	0.03
Cynicism	3.17	3.15	1.44	1.23	-0.38	0.33	0.01
Fear	2.36	2.64	0.98	1.08	0.03	-0.38	-0.23
Sadness	3.26	3.18	1.64	1.47	-0.09	-0.70	-0.01
Surprise	3.85	3.93	1.34	0.96	0.37	-0.03	-0.53
Obsequence	2.95	2.71	1.35	0.93	0.38	0.21	-0.33
Calm	3.13	3.33	1.20	1.12	0.27	0.01	0.71
Funny	2.16	2.44	0.54	0.95	0.06	0.14	0.01

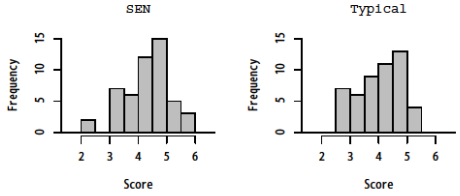
Figure 2: Scatter plot of psychological features.



3.1. Principal component analysis

We applied PCA to the Typical fifty data samples and obtained eigenvectors. We then calculated the principal component (PC) scores of the Typical and SEN data using these eigenvectors. This approach simulates the effect of the SEN data to the Typical data and comparing PC scores is intended to confirm the relative effect of the SEN data. The 1st proportion of variance is 52.2% and the 2nd one is 30.1%. The 1st and the 2nd PCs can sufficiently give an explanation of the psychological space. The right side of table 3 presents the eigenvectors. The 1st and the 2nd PCs seem to express "Valence" and "Activation," which are typical dimensions of emotion, based on the eigenvector scores in Table 3. For instance, in the PC1, the eigenvector score of "anger" is large in the minus direction. The distribution produces a triangular shape and almost all points pertaining to the Typical data gather on each angel of the triangle (Figure 4). The result shows that most of the Typical data are psychologically extreme-featured data and the SEN data fill the low-density regions of the Typical data impression space.

Figure 3: Scatter plot of acoustical features.



3.2. Discussions

This result seems to satisfy the hypothesis pertaining to "(b) filling up low density areas of the existing space," mentioned in chapter 2.1. The SEN data generate more subtle and less overt emotional expressions than the Typical data do (such as "Fear" or "Funny"). Figure 3 shows histograms of the naturalness of each fifty samples of the SEN data and the Typical data. The interval of the histogram is 0.5, from 1.0 to 7.0, and the naturalness of the SEN data appears slightly superior to that of the Typical data. We implement a t test, with a 5% alpha level. The resultant p-value is 0.063, which indicates that there is tendency of significance.

4. ANALYSIS OF ACOUSTICAL FEATURES

To confirm the overview of acoustical features of the SEN and Typical data sets, the left side of table 4 presents the notable acoustical features commonly discussed in emotional speech research. We selected F0 (mean/standard deviation/max/min/range), power (mean/standard deviation/max) and speech duration. In summary statistics, the means and standard deviations of each acoustical feature of the Typical data are greater than those of the SEN data except in some features, such as standard deviation of F0, range of F0 and speech duration.

4.1. Principal component analysis

We also applied PCA to SEN fifty data and Typical fifty data altogether to confirm distributions of each features. The 1st population of variance is 60.8%, the 2nd is 14.3%, and the 3rd is 11.7%. The right side of Table 4 shows eigenvectors. We interpreted the 1st PC as the strength of power and highness of F0 (the direction is opposite), the 2nd PC as a richness of F0 fluctuation (considered from scores of "F0 Std" and "F0 Range") and the 3rd PC as mainly the speech duration. Figure 4 shows a scatter plot of the 1st and 2nd PC for

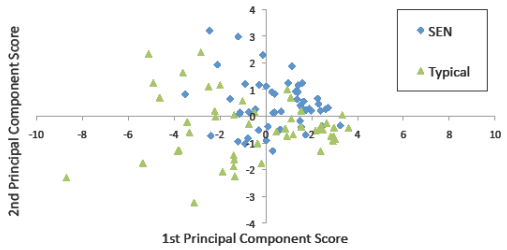
acoustical features. The Typical data set has a large amount of data with very strong power. To focus attention on the bias of the distribution of the 2nd PC, the SEN data set has a tendency to include data with a richness of F0 fluctuation that is comparatively larger than that of the Typical data.

Table 4: Evaluation parameters of acoustical analysis.

Features	Mean		Stddev		Eigenvectors		
	Typical	SEN	Typical	SEN	PC1	PC2	PC3
F0 Mean	372.2	324.8	126.4	87.4	-0.38	-0.22	0.28
F0 Std	67.0	70.5	37.3-	29.2	-0.33	0.52	-0.02
F0 Max	508.7	486.5	194.5	18.0	-0.39	0.13	0.32
F0 Min	225.5	177.1	90.6	65.6	-0.25	-0.57	0.41
F0 Range	284.2	316.0	151.7	111.3	-0.33	0.51	0.14
Power Mean	5.07	3.41	3.38	1.50	-0.38	-0.15	-0.27
Power Std	5.31	3.54	3.38	1.25	-0.37	-0.17	-0.33
Power Max	19.70	13.43	13.08	4.22	-0.37	-0.11	-0.31
Speech Duration	1.956	1.987	0.454	0.387	0.05	0.11	0.59

F0[Hz], Power[dB] Speech Duration[s]

Figure 4: Scatter plot of acoustical features.



4.2. Discussion

The result of PCA indicates the possibility that the richness of variation in the F0 fluctuation generates a detailed expression in the SEN data, which the extreme-featured expression of the Typical data cannot achieve. We implemented a t test with a 5% alpha level, obtaining p-values, for the 1st and 2nd PC, of 0.037 and 0.0013, respectively. This indicates that the SEN data express emotions mainly by F0 fluctuation and the Typical data do so mainly by strength of F0 and Power.

5. RELATIONSHIP BETWEEN PSYCHOLOGICAL AND ACOUSTICAL FEATURES

To verify the relationship between the results of the psychological and acoustical features, we calculated the correlation coefficients of each psycological and acoustical feature (except speech duration) by data type. The amount of an assortment which rejected null hypothesis (decorrelation) was twenty-two in seventy-two pattrens (in the SEN data) and forty in seventy-two patterns (in the Typical data). The mean and the standard deviation scores of the SEN and the Typical data was 0.33/0.09 and 0.47/0.12. Importantly, correlations rarely exist between the psychological and acoustical features in the SEN data, regardless of the tendency of these features described in Chapter 3 and Chapter 4.

Table 5: The ratio of representative BPM types.

Features	BPM Type		
	L%	L%H%	L%HL%
Typical	0.20	0.08	0.72
SEN	0.24	0.18	0.58

Since the emotional expressions in the SEN data are less clear than those in the Typical data in the psychological space, the SEN method might control the acoustical features that are not representative and easy to perceive. Hence, we focused on the tail end of speech, which was considered to influence impressions generally, particularly BPM (boundary pitch movement). We gave J_ToBI annotation on the SEN and Typical data (table 5). We confirmed each ratio of the representative three types of BPM and found there ware large differences between each ratio of the SEN and Typical data. It indicates the possibility that the SEN method can expand psychological diversity by controlling the subtle acoustical features such as BPM.

6. CONCLUSION

In this paper, we propose a new method which involves the use of acting scripts, to collect diverse speech data. We found a possibility that detailed indications by acting scripts generate a psychological diversity that will be described by unrepresentative acoustical features. Future works are: more detailed confirmation of acoustical features concerning BPM, an addition of actors, and a comparison with spontaneous speech.

7. REFERENCES

[1] Campbell, N. 2005. Developments in corpus-based speech synthesis: Approaching natural conversational speech. *IEICE Trans. on Info. & Syst.* E88-D(3), 376-383.

[2] Erickson, D. 2005. Expressive speech: Production, perception and application to speech synthesis. *Acoust. Sci. & Tech.* 26(4), 317-325.

[3] Moriyama, T., Saito, H., Ozawa, S. 2005. Evaluation of the relation between emotional concepts and emotional parameters in speech. *IEICE Trans. on Info. & Sys.* J82-D2(4), 703-711.

[4] Ortony, A., Turner, T.J. 1990. What's basic about basic emotions? *Psychological Review* 97(3), 315-331.

[5] Steidl, S. 2009. *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech*. Logos Verlag Berlin.