Method for Collection of Acted Speech Using Various Situation Scripts

Takahiro Miyajima†, Hideaki Kikuchi†, Katsuhiko Shirai†, Shigeki Okawa‡

†Waseda University
Okubo 3-4-1, Shinjyuku, Tokyo, Japan

‡Chiba Institute of Technology
Tsudanuma 2-17-1, Narashino, Chiba, Japan

E-mail: miyajima@toki.waseda.jp

Abstract

This study was carried out to improve the quality of acted emotional speech. In the recent paradigm shift in speech collection techniques, methods for the collection of high-quality and spontaneous speech has been strongly focused on. However, such methods involve various constraints: such as the difficulty in controlling utterances and sound quality. Hence, our study daringly focuses on acted speech because of its high operability. In this paper, we propose a new method for speech collection by refining acting scripts. We compared the speech collected using our proposed method and that collected using an imitation of the legacy method that was implemented with traditional basic emotional words. The results show the advantage of our proposed method, i.e., the possibility of the generating high F0 fluctuations in acoustical expressions, which is one of the important features of the expressive speech, while ensuring that there is no decline in the naturalness and other psychological features.

Keywords: Emotional Speech, Acted Speech, Acting Script

1. Introduction

A paradigm shift involving "construction and implementation of a general-purpose research corpus," is taking place. This concept arose from the understanding that it is critical to apply experimental results to real-life situations.

Several recent studies on general-purpose emotional corpora (Campbell, 2005; Steidl, 2009) have suggested new approaches to spontaneous speech collection: All speech expressions generated in daily life or in imitations of daily life over a period of several days to years are recorded. The essential conditions for a general-purpose corpus are that the data must be gathered from many people and various utterances in many contexts, must have high spontaneity and naturalness, and must have diverse expression. However, some problems such as the difficulty in controlling the power level and setting of microphones during the recordings persist in the abovementioned mainstream methods.

Therefore, to control the essential conditions with ease and solve the problems, we focus on the potential of acted speech. Though we have a complete understanding of the importance of recording spontaneous day-to-day speech, we suggest a new method that uses professional voice actors and refined acting scripts. In this study, as the first step toward archiving our aim (Figure 1), we confirm the effect of our acting scripts, which are designed to bring about changes in the various features of acted speech.

To this end, we should resolve the issues that most acted speech is not natural and that the psychological and acoustical features of such speech are biased (i.e., seems too artificial). Resolving both issues are important for a general-purpose emotional corpus to be effective across research domains, such as emotional speech synthesis and recognition.

2. Approach

In this study we collected and compared two types of acted speech: (a) speech collected by using our method, and (b) speech collected by using an imitation of legacy methods. In (a), we specified detailed information that described the assumptions for several circumstances of speakers and listeners. We define speech collected via this procedure as "SEN (meaning "thousand" in Japanese) speech data (=SEN data)." In (b), by referring to previous studies (Ortony and Turner, 1990), we adopted some basic emotional words (Table 1) as simple scripts for actors. We define speech data collected by this procedure as "Typical speech data (=Typical data)."

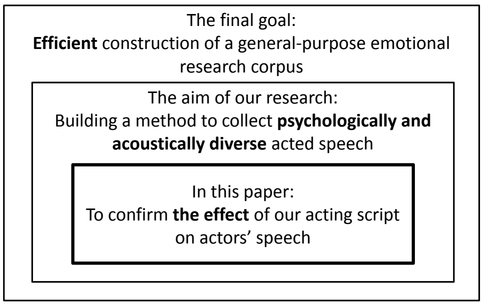
The final goal:
Efficient construction of a general-purpose emotional research corpus
The aim of our research:
Building a method to collect **psychologically and acoustically diverse** acted speech
In this paper:
To confirm **the effect** of our acting script on actors' speech

Figure 1: Purposes of various reaserch steps

Table 1: Summary of basic emotions(Ortony and Turner, 1990)

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

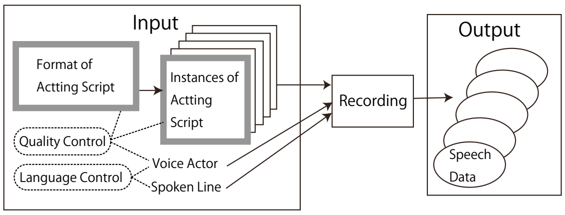

Figure 2: Concept of our proposed method

3. Data Collection

In this chapter, we describe the sequence involved in our method. The concept is shown in Figure 2. We have collected over 2500 speech data samples using our acting scripts by changing detailed factors (table 2). In this paper, we present the procedure for collecting acted speech labeled as "pattern (1)" in table 2, which is prototype speech data collected by using our method.

3.1. Format and Instances of Acting Script

We considered the type of information required by an actor for producing diverse and natural expressions of speech and referred to such information as "format of acting script." We discussed various items (such as personality, context, and situation) and selected a number of significant items after consulting a professional voice actor so as to increase the qualities of acted speech.

To ensure concrete script content (instances of acting scripts), we selected various expressions from TV dialogues for the prototype. To make our selection, we recorded TV programs from a Japanese broadcasting station for 24 h. We selected 304 expressions and produced the same number of acting scripts from 24 h of recorded data, with the aim of generating a high quality of speech in our scripts. Table 3 presents the formats we adopted and instances of each format.

3.2. Recording

3.2.1. Spoken Line and Actor

We then decided upon the actual lines to be spoken. In this prototype, we selected "/aH, so'Hdesuka/" (meaning "Oh, I see" in English) as the spoken line. The reasons for doing so 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 selected line is neutral and will not influence the evaluation of its expression.

3. It includes an accented phase that enables the expression to be varied more easily.

4. It has a high frequency of occurrence in day-to-day conversations. Further, we recruited a professional voice actress with eighteen years of experience in voice acting and theatrical performance.

3.2.2. Recording Procedure

We then extracted one hundred acting scripts wherein the gender of the speaker is female, and eliminated scripts in which the speaker was too young. In this iteration, we collected hundred SEN data and 160 (twice as many as 80 emotional words) Typical data. Typical data are collected by presentation of simple basic emotional words mentioned in Chapter 2.

4. Evaluation and Discussions

Though we obtained many results using collected speech data, we present some representative results in this chapter.

4.1. Overivew of Psychological Features

We compare the psychological features of SEN data and Typical data. To do that, we adopted "emotion express words (Moriyama and Ozawa, 1999)," which consist of nine emotional words (Anger, pleasure, cynicism, fear, sadness, surprise, obsequence, calm, and funny) drawn from forty-six emotional words in various existing works, and the word "naturalness" as a scale for evaluation. Six male and six female university students served as the evaluators. We picked a total of 100 speech data samples, 50 from SEN data and another 50 from Typical data. The evaluation scale was as follows: 1 implied not included at all, 4 implied neutral, and 7 implied included very much.

Then, we applied principal component analysis (PCA) to the selected SEN and Typical data. The 1st proportion of variance is 52.2% and the 2nd proportion is 30.1%. The 1st and the 2nd principal components (PCs) can sufficiently explain the overview of the psychological features. The 1st and the 2nd PCs seem to express "valence" and "activation," which are the typical dimensions of emotion. Figure 3 shows the scatter plot and additional information of the 1st and 2nd PC scores. Thus, it seems that there is rarely a difference between the degree of distribution of SEN and Typical data. These results indicate that the psychological diversity of SEN data will be the almost same as that of Typical data. However, there is a little bias of distribution in the 1st and 2nd quadrant (high activation area), both the SEN and Typical samples. In the score of X-axis (the 1st PC) of the 1st and 2nd quadrant data, the p-value of a t test is 0.38 (no significance, with a 5% alpha level) and that of a F test is 0.03 (heteroscedastic, with a 5% alpha level).

4.2. Evaluation of Naturalness

Figure 4 shows histograms of the results of naturalness evaluation. The interval of the histogram is 0.5, from 1.0 to 7.0, and the naturalness of SEN data appears to be slightly superior to that of 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. Moreover, the theoretical normal distribution curve in the histogram of SEN data seems to be more suitable to real frequency distribution than that of Typical data.

4.3. Overview of Acoustical Features

We selected F0 (mean/standard deviation/max/min/range), power (mean/standard deviation/max) and speech duration as the comprehensive acoustical features. We also applied

Table 2: Various patterns of our acting scripts

Format of Acting Script	Procedure of Building the Instances of Acting Scripts	Spoken Line	Number of Scripts/ Speech Data / Actor
Pattern(1)	Extracted from Expressions in a 24-h TV program	/aH, so'Hdesuka/ (Oh, I See)	304/100/5
Pattern(2)	Extracted From Pattern (1) and a few emotional words are added	/aH, so'Hdesuka/ (Oh, I See)	1400/1400/5
Pattern(3)	Alteration of Pattern (1) based on changes in the personalities of speaker	/yoroshikuonegaishimasu/ (I beg to your kindness)	1000/1000/2

Table 3: Items and examples of acting scripts

Headings		Value of Instance	Example of an Instance
Common	Location / Situation	A Conte, a news show, a variety show, etc. In a school, in a kitchen, in a hospital, etc.	Drama, dormitory hall
	Relationship	School friends, senior and junior entertainers, a reporter and a general, a student and a teacher, father and a daughter, brothers, etc.	Same school and dormitory, good friends
	Context / Background	Apologizing for a past mistake, Evading pursuit of his/her privacy, Talking indifferently to parents' enemies floating tears on his/ her eyes, etc.	Amazed at surrounding commotion
Listeners	Age / Gender	Ten to seventy, male and female	Seventeen or eighteen male
	Career	A student, a MC, a flight attendant, an entertainer, an announcer, a teacher, a doctor, etc.	High school student
	Character	Cool-headed and greedy, brisk and masculine, cheerful and noisy, having his own theory, etc.	Noisy, cluttered
Speakers	Age / Gender	Ten to seventy, male and female	Seventeen, female
	Career	A student, an actress, an announcer, a housewife, a landlady, an intellectual, a commentator, etc.	High School Student
	Character	Sporty, easily elated, neat and clean, strong-minded, brisk and masculine, composed, etc.	Pretending to be a man, ataraxia, insensitive to love

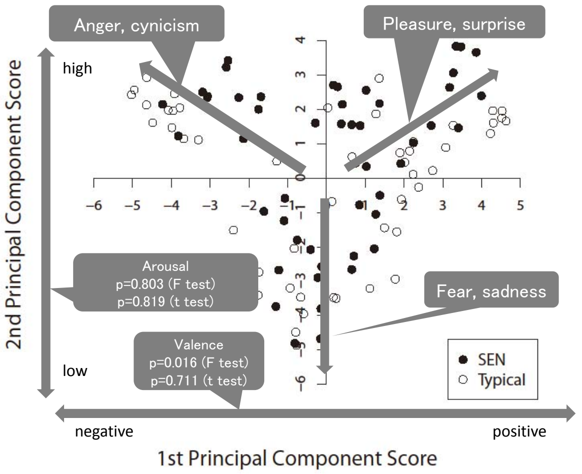

Figure 3: PCA results of pyschological features

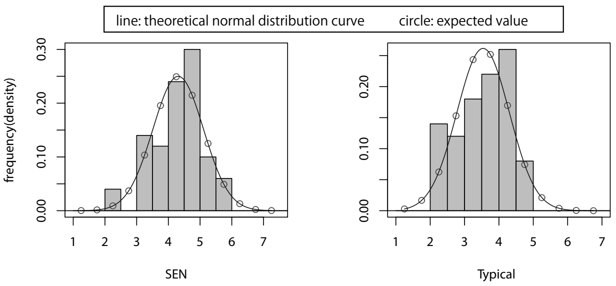

Figure 4: Evaluation results of naturalness

PCA to 50 SEN data and 50 Typical data to confirm the distribution of each feature. The 1st population of variance is 60.8%, 2nd is 14.3%, and 3rd is 11.7%. We interpreted the 1st PC as the strength of power and highness of F0 (the direction is opposite), the 2nd PC as the richness of F0, and the 3rd PC as mainly the speech duration. Figure 5 shows a scatter plot and additional information of the 1st and 2nd PC Scores for acoustical features. Typical data set has a large amount of data with very strong power and high maximum value of F0. To focus attention on the bias of the distribution in the 2nd PC, the tendency of the SEN data-set to include data rich in F0 fluctuation is comparatively larger than that of the Typical data-set.

From the PCA results, diversity of the acoustical features seems to increase when using our proposed method. To confirm this, we determined the F0 curves of each speech and found there is distinctive difference between SEN and Typical data, especially in the F0 curves of the last phoneme ("a"), which is one of the important factor of expressive speech and Japanese intonation (Figure 6, Figure 7); in Figure 6, the X-axis represents data frame number (duration of each curves is normalized to 0.5[ms], and including unvoiced frame) and the Y-axis represents the semitone that have been translated from original F0 values for normalization. The F0 curves of each Typical data are relatively simple, whereas those of each SEN data are varied. The mean score of each frame (Figure 8) shows this tendency more clearly.

4.4. Relationship Between Psychological and Acoustical Features

To verify the relasionship between the results of the psychological and acoustical features, we calculated the corre-

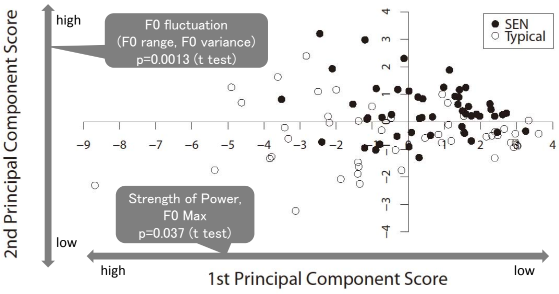

Figure 5: PCA results of acoustical features

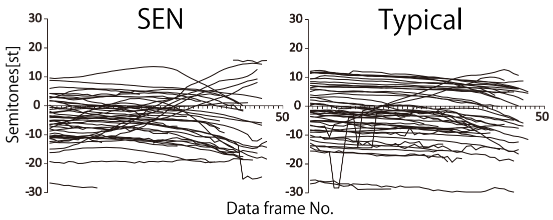

Figure 6: Comparison of normalized F0 curves of the last phoneme “a”

lation coefficients of each psycological and acoustical feature by data type (SEN and Typical data). Figure 9 shows an overview of the results. The scores when each null hypothesis (decorrelation) is rejected are shown and the correlations in the SEN data are smaller than those of the Typical data. Importantly, correlations rarely exist between the psychological and acoustical features in the SEN data, regardless of the tendency of F0 fluctuation mentioned in Chapter 4.3. This could be attributed to the low population of variances of PCA. Because the emotional expressions in the SEN data are less clear than those in the Typical data from the psychological viewpoint, the proposed method might control the acoustical features that are not comprehensive.

5. Conclusion

We propose a new method that involves the use of various situation scripts, for collecting high quality acted speech. The findings in this paper are as follows (these are provided under the condition that acting skill of the actor/actress is high and that they use a specific spoken line):

1. Psychological diversity of SEN data is guaranteed to be almost same as that of Typical data.

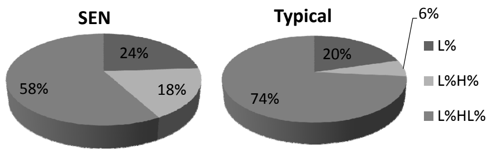

Figure 7: The ratio of representative BPM types of J_ToBI

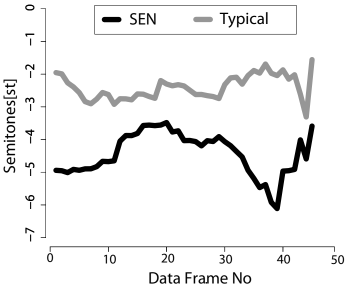

Figure 8: Mean scores of each frame of results in Figure 6

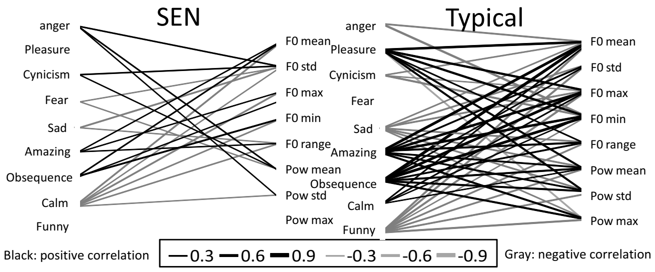

Figure 9: Correlation scores described by the lines and their width

2. The naturalness of SEN data appears superior to that of Typical data.

3. There is an articulate difference between comprehensive acoustical features (especially “strength of power” and “F0 fluctuation”) of SEN and Typical data.

4. In particular, F0 curves of the last phoneme, which is one of the important factors of expressive speech, in SEN data vary while those in Typical data are comparatively flat.

In light of these results, our method provides a new viewpoint for higher availability of acted speech by generating rich F0 fluctuations in speech data without causing a decline in speech naturalness and psychological diversity.

Future works include: To confirm the effects of each item or instance of acting script, more detailed verification of local acoustical features, an addition of actors and spoken lines, preparing various acting scripts, and a comparison of naturalness with spontaneous speech.

6. References

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

H. Moriyama, T. Saito and S. Ozawa. 1999. Evaluation of the relation between emotional concepts and emotional parameters in speech. *IEICE Trans. on Info. & Sys.*, J82–D2(4):703–711.

A. Ortony and T.J. Turner. 1990. What’s the basic about basic emotions? *Psychological Review*, 97(3):315–331.

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