Analysis of laughter events in real science classes by using multiple environment sensor data

Carlos Ishi, Hiroaki Hatano, Norihiro Hagita

ATR Intelligent Robotics and Communication Labs.

carlos@atr.jp, hatano.hiroaki@atr.jp, hagita@atr.jp

10.21437/Interspeech.2014-271

Abstract

The extraction of sound events in environments where a large number of people are present is a challenging problem. In order to tackle that problem, we have been developing a sound environment intelligence system which is able to get information about who is talking, where and when, based on integration of multiple microphone arrays and human tracking technologies. We installed the developed system in a science room of an elementary school, and collected data of real science classes during a period of one month. In the present paper, among the sound activities appearing in the science classes, we focused on the analysis of laughter events, considering that laughter conveys important social functions in communication. Laughter events were extracted by making use of visual displays of spatial-temporal information provided by the developed system. Subjective evaluation of the laughter events revealed relationship between the laughter type (including production, style, and vowel-quality aspects), the functions in communication, and the appropriateness in the classroom context.

Index Terms: laughter, sound activity, non-verbal information, natural conversation, real environment

1. Introduction

Laughter commonly occurs in daily interactions, and is not only simply related to funny situations, but also for expressing some type of attitude, having important social functions in communication. Several works have investigated the functions of laughter and the relationship with acoustic features. For example, it is reported that duration, energy and voicing/unvoicing features change between positive and negative laughter, in a French hospital call center telephone speech [1]. In [2], it is reported that the first formant is raised and vowels are centralized (schwa), by analyzing English acted laughter data of several speakers. In [3-4], it is reported that mirthful laughter and polite laughter differ in terms of duration, the number of calls (syllables), pitch and spectral shapes, in Japanese telephone conversational dialogue speech.

However, most of works related to laughter (as the ones cited above) use dialogue data of pair of speakers in favorable acoustic conditions, while only a few deals with situations where multiple speakers are present in the environment. The main reason is that the extraction of sound events in environments where a large number of people are present is a difficult and challenging problem.

On one hand, microphone array processing technologies have been applied for localization of sounds by "intelligent rooms" (e.g. [5-8]). In order to deal with situations where multiple speakers talk and move, we have been developing a sound environment intelligence system which is able to get information about who is talking, where and when, based on integration of multiple microphone arrays and human tracking technologies [7-9].

We obtained permission an elementary school to install the developed system in its science classroom, and collected data of real science classes during a period of one month. Among the several sound events appearing in the classroom, in the present paper, we present analysis result for laughter events appearing during the science classes.

Thus, in the present work, we make use of the spatial-temporal information extracted from the sound environment intelligence system, and conduct analyses on the laughter events by focusing on the following questions: 1) What are the types of laughter occurring during the classes? 2) Is there a relationship between the laughter types and the communicative functions conveyed by them? 3) Is a laughter appropriate or not to the class context?

2. Analysis data

2.1. Description of the data collected in the science classroom

Fig. 1 shows a picture of the science room where data was collected. There are 8 desks for the science experiments from which the front 6 ones are actually used. Thus, we installed 6 microphone arrays over each of the 6 desks. After discussion with the school side, we were allowed to locate the arrays at about 2 meters height right over the desk's sink, so that they do not obstruct the student's field of view and the teachers do not hit their heads. (See Fig. 1.) For the human position detection, multiple Kinect sensors were attached in the ceilings.

Figure 1. *The science classroom where the developed system was installed.*

Fig. 2 shows the geometry of the microphone array used for audio data collection. Sixteen microphones are distributed in the 3D axes (approximately over a semi-sphere of 30 cm diameter) for allowing sound direction estimation in the 3D space (i.e., estimation of azimuth and elevation angles). An array frame was designed to fix sixteen silicon microphones according to the geometry of Fig. 2. As the multi-channel

audio capture device, we used the TD-BD16ADUSB from Tokyo Electron Devices Ltda.

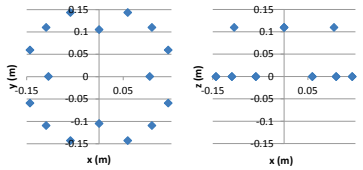

Figure 2. *The geometry of the microphone array.*

For sound direction estimation by each microphone array, we used our implemented system based on the MUSIC (MUltiple SIgnal Classification) method, which is able to provide 1 degree angle resolution in the 3D space and 0.1 second time resolution in real-time, by a 2GHz CPU [10].

For the human tracking, we used a particle filter-based 3D human localization using the multiple Kinect sensor data [9]. The human position data is obtained in 33 ~ 66 ms resolution. Although a 2D human tracking based on LRF (Laser Range Finders) is another alternative [7], we opted for using the Kinects, due to the large number of students and the appropriateness for attaching sensors in the ceiling.

Multi-modal data was collected during a period of one month (in Feb. 2013) for the science classes of the 5th year. The 5th year is composed by around 120 students divided in four classes (around 30 students per class). Thus, the number of students per class is around 30 (5 to 6 students per desk) and the number of teachers is two (one is the class teacher, while the other is the science teacher).

Fig. 3 shows two examples of sound direction estimation results by the 6 arrays in the science classroom. Dashed lines are drawn every meter. The straight lines from each array are the detected sound directions. The colors of the lines represent different heights every half a meter. (Heights are able to be represented thanks to the 3D estimation of sound directions). The circles (other than the array positions) represent the detected human positions by the set of Kinect sensors. It can be observed that the locations of about five students are detected around each array. The example in the left shows an instant where the teacher is doing explanation in front of the classroom, while the example in the right shows an instant where two students are simultaneously speaking.

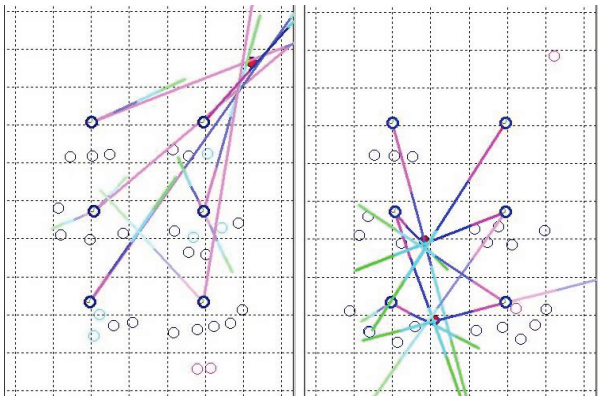

Figure 3. *Examples of sound directions detected by the six microphone arrays in the science classroom.*

Fig. 4 shows example of sound directions detected by the six microphone arrays in the science classroom, for an interval of 20 seconds, where students are interacting with the teacher or with each other during the experiments. The vertical axis represents azimuth angles (-180 to 180 degrees), while the different colors represent different elevation angle ranges (the activities of the students around the desks correspond to elevation angles represented by pink). Segments of sound activities around each of the six desks can be identified in several directions in the displays of the sound direction results.

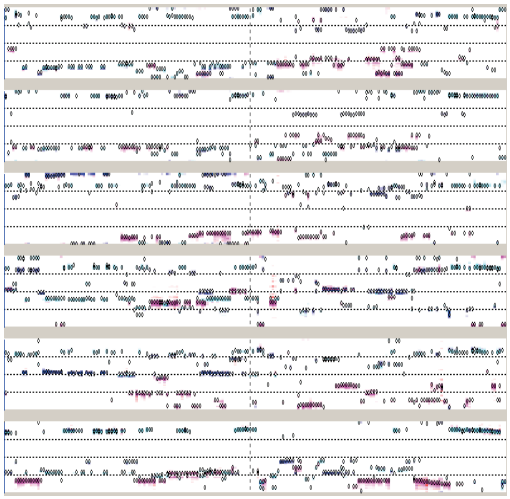

Figure 4. *Example of sound directions detected by the six microphone arrays in the science classroom for an interval of 20 seconds, where students are interacting with the teacher during the experiments. Vertical axis is azimuth angle from -180 to 180 degrees (dashed lines each 90 degrees), horizontal axis is time.*

Data of 5 days (including a total of 10 hour-classes) were used in the laughter analysis. Laughs often occur simultaneously with other laughs or speech utterances, so that a sufficient acoustic quality is difficult to be achieved even after sound separation based on microphone array processing. For the present analysis, the laughter events were manually identified and segmented (by a research assistant).

The laughter events in each of the 6 desks were segmented by making use of the spatial information provided by the sound direction estimation results. During the segmentation process, the annotator was able to look at the spatial and temporal displays similar to the ones shown in Figs. 3 and 4, and the spectrogram displays for each microphone array. By visually identifying the placements of the sound sources, the annotator was also able to select the microphone array signal closest to the target sound source and listen to specified intervals and segment the laughter portions.

762 laughter events were extracted from the database. 0f these, about 10% were from teachers while the rest were from students. We first observed that the amount of laughter was related to the contents of the classes. Fig. 5 shows the distributions of the laugh events for the ten classes. The global theme of all ten classes was regarding “the birth of life”. The class group number (5-1 ~ 5-4) and the class style are specified in the vertical axis of Fig. 5. The duration of the short videos were about 5 minutes.

In Fig. 5, a high number of occurrences of laughter (more than 100) can be observed in the classes regarding “weight experience”. In these classes, the students were asked to wear

a weighted cloth simulating the weight of a baby in the abdomen during several steps in the pregnancy period. This was the reason for the high occurrence of laugh events during the classes.

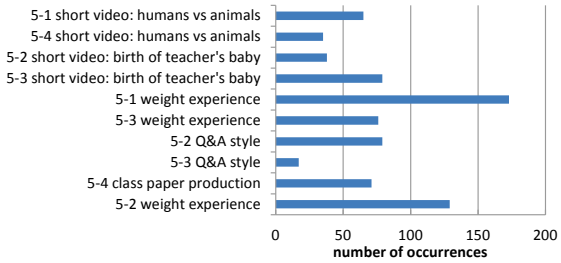

Figure 5. *Distribution of the laugh events for ten science classes.*

The laughter events were manually classified according to single or multiple laughs per desk (i.e., multiple speakers around the same desk laughing at the same time). This is to account for the fact that multiple laughs around the same desk are more difficult to be separated so that for subsequent annotation of laughing types it would be difficult to be sure which of laughter event the annotation is attributed to. Multiple laughs around the same desk occurred in 8% (59/762) of the laughter events. These were removed from the analysis in the present work.

For subsequent analyses, the signal from the array closest to the speaker was selected, based on the displays of the spatial information from the sound direction estimations (as in the examples shown in Fig. 3) and segmented including five seconds before and after the laughter portion to take context into account.

2.2. Annotation data

We analyzed the laughter events from the following viewpoints: the types of laughter occurring during the classes; relationship between the laughter types and the communicative functions conveyed by them; and the appropriateness of a laughter event in the class context.

In order to account for the above issues, the labels shown in Table 1 were elaborated, by taking into account knowledge from past works on laughter, and preliminary annotation evaluation. The first three categories are related to the laughter manner, while the last two categories are related to the functionalities and appropriateness of the laughter.

Table 1. *List of the annotation labels accounting different aspects of laughter events. (The terms in parenthesis are the original Japanese terms used in the annotation.)*

1.	***Laughter production type***: {breathiness over the whole laughter segment ("kisoku"), alternated pattern of breathy and non-breathy parts ("iki ari to iki nashi no kougo"), relaxed (vocal folds relaxed, absence of breathiness: "shikan"), laughter during inhalation ("hikiwarai")}
2.	***Laughter style***: {secretly ("hisohiso"), giggle/chuckle ("kusukusu"), guffaw ("geragera"), sneer ("hanawarai")}
3.	***Vowel-quality of the laughter***: {"hahaha", "hehehe", "hihihi", "hohoho", "huhuhu", schwa (central vowel)}
4.	***Laughter function***: {funny/amused/joy/mirthful laugh ("omoshiroi", "okashii", "tanoshii"), social/polite laugh ("aisowarai"), bitter/embarrassed laugh ("nigawarai"), self-conscious laugh ("terewarai"), inviting laugh ("sasoiwarai"), contagious laugh ("tsurarewarai", "moraiwarai"), depreciatory/derision laugh ("mikudashiwarai"), dumbfounded laugh ("akirewarai"), untrue laugh ("usowarai")}
5.	***Relationship with the class***: {related, partly related (for example when the teacher makes a question related to the class contents, but the student answer with a joke which is not related to the class contents), not related}

Three native speakers of Japanese (research assistants) annotated the labels in Table 1, by listening to the segmented intervals (including five seconds before and after the laughter portions.) For the label items in "laughter style" and "laughter functions" (items 2 and 4 in Table 1), annotators were allowed to select more than one item per laughter event. No specific constraints were imposed for the number of times for listening, or the order for annotating all items in Table 1. One of the annotators felt more comfortable to conduct the items 1, 2, 3 and 5 at the same time first, and the laughter functions in item 4 later.

The agreement rates (in terms of kappa coefficients) among the labels of the 3 annotators were moderate for "vowel quality" (0.44, 0.56, 0.44), and "laughter functions" (0.45, 0.54, 0.46), fair for "laughter style" (0.33, 0.47, 0.31), and good for "laughter production type" (0.57, 0.71, 0.54), and "relationship with the class contents" (0.59, 0.69, 0.58). The agreement rates in "laughter style" are low because the number of tokens where specific laughter styles appeared was small. The laughter events where 2 or more annotators agreed were used for the subsequent analyses.

The number of laughter calls (individual syllables in an /h/-vowel sequence) was also annotated for each laughter event, by looking at the spectrogram displays.

3. Analysis of the laughter events in the science classroom

3.1. Analysis of laughter types

The relationships between production type, style, and vowel quality were analyzed. In the present paper we show the results for the most relevant relationships found. Fig. 6 shows the distributions of the production type categories for the different categories of laughter style.

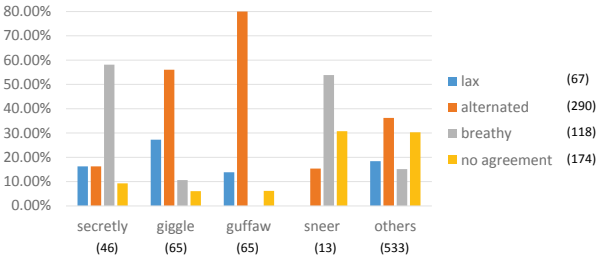

Figure 6. *Distribution of the production type categories for the different categories of laughter style. The total number of utterances is shown within brackets.*

It can be observed in Fig. 6 that "alternated" patterns of breathy and non-breathy parts are predominant in "giggle" and "guffaw", while "breathy" laughs are predominant in sneer and "secretly" laughter.

No clear relationship was found between the vowel quality and the production type or laughter style.

3.2. Analysis of the laughter types and the laughter functions

Fig. 7 shows the distributions of the production type categories (top panel), the vowel-quality categories (mid panel) and the number of laughter calls (bottom panel), for the different categories of laughter functions. Note that the laughter function categories contain single or multiple items, as a result of multiple selections during the annotation. The category "negative" includes "depreciatory laughs" and "dumbfounded laughs".

It can be observed in the top panel of Fig. 7 that "alternated" patterns of breathy and non-breathy parts are predominant in "funny"-only laughs, and decrease when "social/bitter" laughs co-occur. On the other hand, it can be noted that the occurrence rate of "breathy" patterns increase in "social/bitter" laughs.

Regarding the vowel-quality, it can be observed in the mid panel of Fig. 7 that schwa (central) vowels are predominant in all categories. However, it can be noted that "hahaha" also appear with high frequency in "funny"-only laughs, while "huhuhu" appears with high frequency in "funny + social/bitter" laughs.

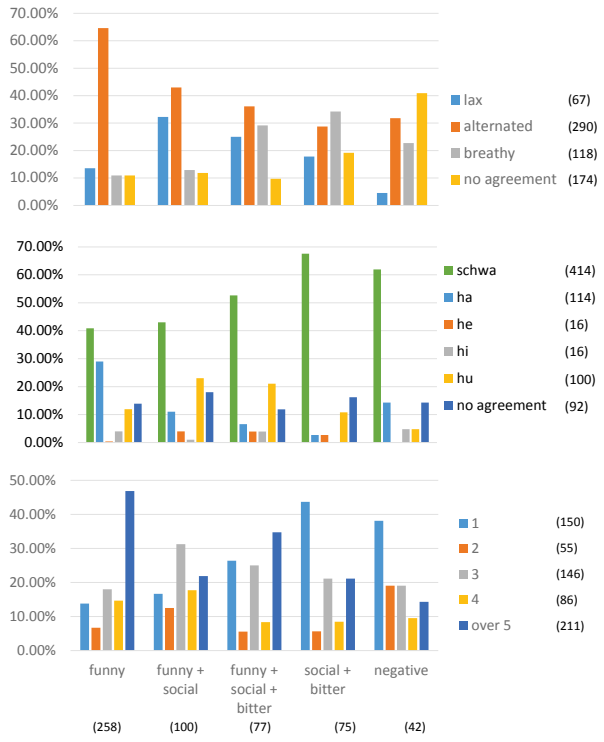

Figure 7. *Distributions of the production type categories (top panel), the vowel-quality categories (mid panel) and the number of laughter calls (bottom panel), for the different categories of laughter functions. The total number of utterances is shown within brackets.*

Regarding the number of laughter calls (individual syllables in an /h/-vowel sequence) in the bottom panel of Fig. 7, it can be observed that "funny" laughs tend to have predominance of more than 5 laughter calls per event, while "social/bitter" (without "funny" expression) and "negative" laughs have predominance of only 1 laughter call.

3.3. Analysis of the laughter types and the laughter situations

Fig. 8 shows the distributions of the laughter styles according to its relationship with the class contents.

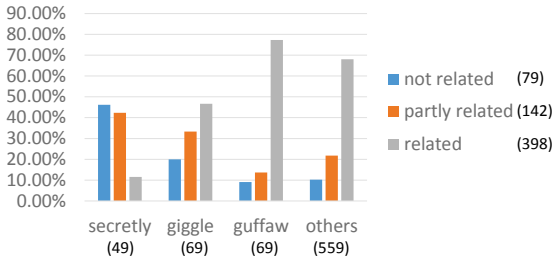

Figure 8: *Distributions of the laughter style for the different categories of relationship with the class contents. The total number of utterances is shown within brackets.*

It can be observed in Fig. 8 that "secretly" laughs have predominance of laughter events which are "not related" to the class contents, while "guffaw" and "other" laughter styles are predominantly "related" to the class contents. From these results one can say that the 5th year elementary school students are already able to (partly) concern about the laughter style according to the situation in the classroom context.

4. Conclusions

In the present work, we analyzed laughter events appearing in real science classes recorded in an elementary school. The identification of laughter events in a large environment (classroom) with a large number of people (about thirty students and two teachers) was made possible thanks to the use of spatial-temporal information, provided by multiple microphone arrays and human tracking technologies.

Analysis results of laughter events in the science classroom revealed relationships between the laughter type, the laughter function and the relationship with the class contents. "Funny"-only laughter (which can be considered as spontaneous laughter) tends to have high occurrence of alternated breathy and non-breathy patterns, "ha" vowel quality, and large number of calls (syllables), while "social/bitter"-only laughter (which can be considered as non-spontaneous laughter) tends to have high occurrence of breathy patterns and single calls. "Secret" laughter tends to be breathy, and to appear in situations where the laughter is not related to the class contents.

Future works include evaluation of acoustic features for automatic detection of laughter and other relevant sound events from the multi-modal data. We also plan to install our developed system in home environments and analyze daily-life data. A comparison between laughter events of adults and children would also be an interesting future topic.

5. Acknowledgements

This study was supported by JSPS KAKENHI Grant (No. 21118003, 21118008 and 25240042). We thank all the staff and students of the "Kyoto-fu Seikacho Higashi-Hikari" Elementary School for collaborating in the data collection. We also thank all research assistants of the ATR group who contributed with the annotations and data analyses.

6. References

[1] Devillers, L. & Vidrascu, L., "Positive and negative emotional states behind the laughs in spontaneous spoken dialogs", Proc. of Interdisciplinary Workshop on The Phonetics of Laughter, 37-40, 2007.

[2] Szameitat, D. P., Darwin, C. J., Szameitat, A. J., Wildgruber, D., & Alter, K. Formant characteristics of human laughter. J Voice, 25, 32-37, 2011.

[3] Campbell, N., "Whom we laugh with affects how we laugh", Proc. of Interdisciplinary Workshop on The Phonetics of Laughter, 61-65, 2007.

[4] Tanaka, H. & Campbell, N., "Acoustic features of four types of laughter in natural conversational speech", Proc. of ICPhS XVII, 1958-1961, 2011.

[5] Y. Sasaki, S. Kagami, H. Mizoguchi, T. Enomoto "A predefined command recognition system using a ceiling microphone array in noisy housing environments," in *Proc. of IROS 2008*, Nice, France, 2008, pp. 2178–2184.

[6] R. Chakraborty, C. Nadeu, T. Butko, "Detection and positioning of overlapped sounds in a room environment," in *Proc. of Interspeech 2012*, Portland, USA, 2012.

[7] J. Even, C. T. Ishi, P. Heracleous, T. Miyashita, N. Hagita: "Combining laser range finders and local steered response power for audio monitoring," Proc. IROS 2012: 986-991, 2012.

[8] C. Ishi, J. Even, N. Hagita, "Using multiple microphone arrays and reflections for 3D localization of sound sources," in *Proc. of IROS 2013*, Tokyo, Japan, 2013

[9] D. Brscic, T. Kanda, T. Ikeda, T. Miyashita, "Person tracking in large public spaces using 3D range sensors", *IEEE Transactions on Human-Machine Systems*, pp. 522-534, 2013.

[10] C. T. Ishi, O. Chatot, H. Ishiguro, N. Hagita, "Evaluation of a MUSIC-based real-time sound localization of multiple sound sources in real noisy environments," in *Proc. of the 2009 IEEE/RSJ Intl. Conf. on Intelligent Robots and System*, St. Louis, USA, 2009, pp. 2027–2032.