Information of corresponding author, Hiroki Tanaka.

TEL: +81-743-72-5265

FAX: +81-743-72-5269

E-mail: hiroki-tan@is.naist.jp

Classification of Social Laughter in Natural Conversational Speech

Hiroki Tanaka [a,*], *Nick Campbell* [a,b]

[a] *Augmented Human Communication Laboratory, Nara Institute of Science and Technology, Takayama-cho 8916-5, Ikoma-shi, Nara, Japan*
[b] *Speech Communication Lab, CLCS, Trinity College Dublin, College Green, Dublin, Ireland*

Abstract

We report progress towards developing a sensor module that categorizes types of laughter for application in dialogue systems or social-skills training situations. The module will also function as a component to measure discourse engagement in natural conversational speech. This paper presents the results of an analysis into the sounds of human laughter in a very large corpus of naturally-occurring conversational speech and our classification of the laughter types according to social function. Various types of laughter were categorized into either polite or genuinely mirthful categories and the analysis of these laughs forms the core of this report. Statistical analysis of the acoustic features of each laugh was performed and a principal component analysis and classification tree analysis were performed to determine the main contributing factors in each case. A statistical model was then trained

using a support vector machine to predict the most likely category for each laugh in both speaker-specific and speaker-independent manner. Better than 70% accuracy was obtained in automatic classification tests.

Keywords: Laughter, Prosody, Paralinguistic information, Non-verbal behaviour, Classification, Support vector machines

1. Introduction

In human-human interaction, communication involves both verbal and nonverbal information, and the latter serves especially to express discourse engagement. One of the most common nonverbal vocalizations in social conversation is laughter [1]. which is also reported as the most frequently annotated acoustic nonverbal behavior in meeting corpora [2] where 8.6% of the time a person vocalizes in a meeting is spent on laughing and 0.8% is spent on laughing while talking. Laughter is a universal and prominent feature of human communication [3], and expressed by both vocal and facial expressions. It is a powerful affective and social signal [4]. There is no culture where laughter is not found. However, current dialogue systems and computer-based social skills training (a training method for people with autism or asperger syndrome to learn social function [5]) do not take into

account laughter [6].

In a seminal study of the segmentation of laughs, Trouvain [7] suggests that we consider laughter as articulated speech, where at the low level there are sound segments that are either vowels or consonants. At the next higher level, there are syllables consisting of sound segments. The next higher level deals with larger units such as phrases which are made up of several syllables. Owren [8] recommends the term 'bout' for the longer sequence, and 'call' for the individual syllables; we will adopt that terminology in this study.

Some earlier work on the automatic segmentation of laughter has been reported in the literature. Khiet P. Truong *et al.* [9] reported automatic laughter segmentation in meetings. They performed laughter vs speech discrimination experiments comparing traditional spectral features and acoustic phonetic features, and concluded that the performance of laughter segmentation can be improved by incorporating phonetic knowledge into the models. Scherer *et al.* [10] reported that the total accuracy of detecting laughter from natural discourse in human-computer interaction reached over 90% in online and offline detection experiments with speech and visual information. Kennedy and Ellis [11] focused on joint laughter in meetings, which means participants (more than just one) laugh simultaneously [12, 13, 14], and they

obtained detection results with a correct accept rate of 87% and a false alarm rate of 13% by using Support Vector Machines.

Types of laughter vary in natural conversational speech, and some classifications have been reported in the literature regarding different categories of laughter. Most types of laughter were discussed in [15], and the major work is the discrimination of laughter into two types, voiced and unvoiced, based on acoustics [16, 17]. Laurence *et al.* [18] deal with a study of laughs in spontaneous speech and explore the positive and negative valence of laughter towards their global aim of detecting emotional behavior in speech. The conclusion of their acoustic analysis is that unvoiced laughs are more often perceived as negative and voiced segments as positive. Previous work in the literature has also discussed whether laughter patterns can be defined through stereotypes [19, 7, 20]. However, laughter is not simply positive or negative, or even defined by stereotypes; it is quite usual for people to infer different degrees of emotion and engagement based on its perceptions, and it is common for people to make use of social laughter in sophisticated social interaction. In this study we tested perceptual types of laughter to determine the main characteristics of laughter in social interaction by reference to the above previous studies.

Automatic classification of four phonetic types of laughter in a natural-speech conversation corpus was conducted by Campbell *et al.* [21], based on perceptual impressions of laughter, in which a laughter episode is considered as a sequence of speech-like phonetic segments (after Bachorowski *et al* [19]). The work described 4 different laughter types: voiced, chuckle, breathy and nasal, and modeled each laugh as composed of different combinations of these segments using Hidden Markov Models (HMMs) statistical classification. The study reported an automatic discrimination using 3 to 15 states with mfcc-based HMMs for 4 functions of laughter (hearty, amused, satirical, and polite). In categorizing emotional classification the work achieved 76% accuracy. However because of the hidden nature of the statistical modeling the report did not provide explicit details about which specific acoustic features contributed to the various categorizations of the laughter.

We report progress towards developing a sensor module that categorizes types of laughter for application in dialogue systems or social skills training situations. In the present study we only make use of the audio information but recognize that facial expression also carries an important channel of communicative information [22, 23]. This paper reports a study of laughs in a corpus of human-human dialogues recorded from Japanese telephone conver-

sational speech [24]. We employed a corpus of natural spontaneous speech where laughter occurred naturally as a consequence of the dialogue interaction. We specifically avoid the use of contrived laughter or even specifically elicited laughs since they may not be representative of natural spontaneous interaction.

In the following sections we first provide details of the corpus, then introduce two Experiments. Experiment 1: a perceptual test by Japanese students to determine the number and types of easily discriminated laughter, and Experiment 2: describing the acoustic feature extraction, presenting the results of an analysis of the main acoustic features and finally reporting a classification of type of laughter using statistical methods.

2. Data: Natural Types of Laughter

We used two types of Japanese corpora. First, the Expressive Speech Processing (ESP) corpus [24] was used for this study. The speech data were recorded over a period of several months, with paid volunteers coming to an office building in a large city in Western Japan once a week to talk with specific partners in a separate part of the same building over an office telephone. While talking, they each wore a head-mounted Sennheiser HMD-410

close-talking dynamic microphone and recorded their speech directly to DAT (digital audio tape) at a sampling rate of 48kHz. They did not see their partners or socialize with them outside of the recording sessions. Partner combinations were controlled for sex, age, and familiarity, and all recordings were transcribed and time-aligned for subsequent analysis. Recordings continued for a maximum of eleven sessions between each pair which were numbered consecutively as session 01 to session 11. The additional eleventh session was only used in the case of absence of one of the volunteers from one of the regular sessions but provided useful additional material. Each conversation lasted for a period of thirty minutes. In all, ten people took part as speakers in the recordings, five male and five female. Six were Japanese, two Chinese, and two native speakers of American English. All were resident and working in Japan at the time. The speech data were transferred to a computer and segmented into separate files, each containing a single utterance for manual transcription by professional transcribers. Laughs were marked with a special diacritic, and laughing speech was also bracketed to show which sections of ordinary speech were spoken with a laughing voice. Laughs were transcribed using the Japanese Katakana phonetic orthography, wherever possible, alongside the use of the identifying symbol. The present

analysis focuses on speakers JMA (age 20s) JMB (age 20s), EMA (age 20s), EFA (age 20s), CMA (age 30s), and CFA (age 20s) to confirm that the same types of laughter are common across different native language groups. The other speakers are all female and similar to the speaker FAN in terms of age, sex, and native language, and thus we selected one female speaker as representative for the present analysis. JMC is omitted because his speech data is insufficient. The initial letters J, C and E indicate native speaker of Japanese, Chinese, and English respectively, M or F indicates the gender of speaker, and A or B indicates the session group of speakers as used for a different experiment.

Second, data from speaker FAN (age 30s) was also used in this report. The FAN subset of the ESP corpus was recorded over a period of five years with everyday conversational speech collected from a single female volunteer wearing high-quality head-mounted microphones, recording her speech to a small Mini-Disc recorder as she went about her daily life. This part of the corpus features a lot of speech in various situations and much simple, repetitive and unstructured talk that illustrates how we spontaneously speak in everyday situations. Speaker FAN was a young female Japanese who personally provided more than 600 hours of usable speech material. Because we

were not able to enter into contractual agreements with her various interlocutors, only the voice of FAN herself has been transcribed or analyzed. While this material is less useful for the analysis of conversational interaction, it provides valuable insights into the range of voice qualities and speaking styles used by one person throughout her daily life.

The study reported here includes two perceptual experiments. The first tested for perceptual types of laughter using Japanese students as subjects listening to the natural conversational speech recordings. We used these results to confirm the classification into the most easily perceived classes of laughter in the corpus. The second tested the degree to which opinions were shared between respondents in the initial classification. For both experiments we predicted the following:

1. In social communication, people do not use hearty laughter with high frequency, rather they typically express polite social laughter (Experiment 1);
2. There are some important acoustic features that can be used to distinctively classify the types of laughter; these includes laughter specific parameters such as the number of the calls; and
3. Automatic classification of laughter is possible at rates greater than

chance in both closed and open tests (Experiment 2).

3. Experiment 1

This experiment concerned the annotation of types of laughter found in the ESP corpus and we chose conversations between JMA and JMB, and JMA and EFA as illustrative.

3.1. Method

We recruited 20 Japanese students (age 23 to 26), and they downloaded wav files from three of the thirty-minute sessions (JMA-EFA; session 03, JMA-JMB; session 03, and JMA-JMB; session 11). Male speaker JMA is the common factor here, and we noticed that his utterance and laughter would change depending on the partner information and the number of sessions (i.e., 'familiarity') [25, 26]. Annotators were free to select one from the list of three conversations for annotation, and were required to categorize both JMA ' s and partner ' s laughter. 8 students choose JMA-EFA; session 03, 6 students choose JMA-JMB; session 03, and 6 students choose JMA-JMB; session 11.

We determined types of laugher by reference to previous work [21, 27], as 'mirthful', 'polite', 'derisive', and 'others' because this research utilises

spontaneous speech data, and thus derisive laughter is sometimes included in the corpus [28]. Because hearty laugh and amused laugh in [21] were sometimes difficult to distinguish, these were both included under the category of mirthful laughter. The ESP corpus has been richly transcribed and subjects worked from phonetic laughter transcriptions such as 'hahaha', 'hihihi', or 'huhuhu'.

The instruction page for the annotation exercise was created in html and students carried out annotations following these instruction in their own space, either at home or in the laboratory. The resulting annotation was sent to the first and second author by E-mail.

3.2. Result: Main Types of Laughter

The 20 annotator agreement was measured by Multi Cohen's kappa-coefficient which calculates agreement beyond chance by distinguishing the observed agreement (A_{obs}) from the agreement by chance (A_{ch}), according to the following:

$$\kappa = (A_{obs} - A_{ch}) \,/\, (1 - A_{ch}) \tag{1}$$

We implemented pair-wise kappa for all annotator pairs, and obtained a kappa value 0.46, which corresponds to moderate agreement according

to the scale proposed by [29]. It must be noted that low kappa scores do not necessarily mean low agreement [30]: if the annotators share certain assumptions of the data, their chance agreement is higher, and the above formula gives smaller kappa values.

As a result, we found that mirthful and polite laughs account for 90 percent of all laughs in these samples of human social interaction and only a very small number of derisive laughs were heard. Approximately 8% of the time when a person vocalizes in natural dialogue is spent on laughing (ref. Table 1). The table shows counts of labels both for laughs and laughing speech, though we omit any results for laughing speech from this study because of its linguistic complexity.

<Table 1 around here>

Experiment 1 was carried out to determine which types of laughter were most readily perceived by typical Japanese students, and we confirmed hypothesis 1; In social communication, people do not use hearty laughter with high frequency, rather they typically express polite social laughter. Since people with autism perceive polite laughter as mirthful laughter [31], a sensor module which classifies polite laughs is considered beneficial for social skills training situations. Our research is directed to this goal.

The main types of laughter in these recordings were determined to be polite and mirthful (accounting for 90% of the laughs), and the number of other types of laughter is too small to be integrated into a sensor module reflecting social functions. Thus we take the majority vote of the observers, and categorized two basic types: mirthful laughs henceforth labeled 'm' and polite laughs labeled 'p' for use in Experiment 2.

4. Experiment 2

This experiment concerned an analysis of the acoustic parameters of the two types of laughter we defined above, and was implemented in classification of natural laughs by using Support Vector Machines, a widely-used high-performance statistical classifier.

4.1. Segmentation and Annotation

In Experiment 1 we determined two types of laughter that are common in Japanese social conversation, polite and mirthful. Experiment 2 utilized this result and two small classes (derisive and others) are removed because these is not enough data to use. We explored the variation as the number of speakers was increased. Table 2 shows the number of laughs used for this Experiment. For the analysis of acoustic features we used speakers JMA,

JMB, and FAN, and for the test of cross-prediction by Support Vector Machine the speakers JMB, FAN, EMA, EFA, CMA, and CFA were selected to evaluate the generalization ability of the classifier.

The choice of partner is important in classifying these two types of laughter; in this report the frequent speakers, JMA and JFA, who talk with almost all others were chosen. Thus, we select the following sessions; CMA-JFA, EFA-JFA, EMA-JFA, JMA-JFA, CFA-JMA, CMA-JMA, EFA-JMA, EMA-JMA, and JMB-JMA. The ESP corpus has rich transcription of all utterances and laughter segmentation was performed using linguistic label time-stamp information. An annotator manually labelled each laugh thus excised into either polite or mirthful categories according to the results obtained from Experiment 1.

<Table 2 around here>

4.2. Acoustic Feature Extraction

<Table 3 around here>

The prosodic acoustic features for each laugh were calculated by a software programme we wrote using the Snack speech processing Toolkit, part of the Tcl/Tk programming language [32]. Explicit prosodic features were included for analysis in this report because our earlier work had used mfcc

parameters only. Overall classification accuracy from the mfcc alone is less than that obtained when using higher-level prosodic features such as F0, amplitude, duration, and their derivatives. In addition to these fundamental prosodic parameters, spectral tilt or shape parameters and positional parameters (fvcd, ppct, and fpct) were estimated to facilitate voice quality descriptions and to encode the acoustic dynamics of the laughter.

The features we tested were measures of fundamental frequency, speech amplitude, and spectral tilt. For fundamental frequency and power, we calculated the mean, maximum, and minimum values measured across each laugh (fmean, fmax, fmin, pmean, pmax, and pmin), as well as the position of the maximum in relative percentage values within each speech waveform (fpct, and ppct). We estimated spectral tilt from the difference between the first harmonic and the amplitude of the third formant (h1a3) after Hansen [33], and by the difference between the first harmonic and the second harmonic (h1h2), as well as taking into account the amplitude of first harmonic (h1) and third formant (a3) respectively. We also measured duration of the laugh (dn) as well as the amount of voicing it contained (fvcd).

We extracted 'No.Call' (The number of calls in a bout) as a further feature for our analysis. The call unit segmentation is implemented by use of an mfcc

3-state Hidden Markov Model with, which achieved over 87% accuracy for each of the four call types within a bout (voiced, ingressive, chuckle, and nasal) as reported in [34]. We calculated the correlation coefficient between duration and No.calls of JMA and obtained a correlation of 0.91 ($p < 0.001$ (signif)). Although highly correlated we consider the number of calls to be a relevant parameter in our modeling as it may distinguish between many short calls and few longer ones each having the same overall bout duration. Actually, approximately 1 % of accuracy rate is changed according to the inclusion each of these features against each speaker in our pilot experiment.

Two further dynamic parameters 'F0moveAB' and 'F0moveAN' were also extracted. As Figure 1 shows, these parameters need $F0avg2a$ which represents average logarithm of pitch within a first (A) call, and $F0tgt2b$ (Second (B) call) and $F0tgt2n$ (Final (N) call) which represents the pitch target at the end of each call by a simple regression coefficient. Pitch change between the first and the second call (F0moveAB) is calculated $F0avg2a - F0tgt2b$, and that between the first and the final call (F0moveAN) is also calculated $F0avg2a - F0tgt2n$. When there is one call within a bout, we set these dynamic parameters to zero.

<Figure 1 around here>

4.3. Statistical Analysis Tool

This section reports a statistical analysis of human laughter which was annotated as either polite or mirthful, using parameter reduction by means of Principal Component Analysis and Classification Trees. An automatic classification of the two types of laughter is reported in this section. The statistical analyses were performed using the free public-domain software package R [35]. Specifically, we used the additional option package 'tree' for Classification Tree and package 'e1071' for the Support Vector Machine analysis.

4.4. Principal Component Analysis

We split the data into training and test set (JMA; training: 206, test: 61, JMB; training: 191, test: 71, FAN; training: 270, test: 61), and the number of label 'p' and 'm' are balanced in each set. We ensure that the test material does not appear anywhere except in a validation experiment.

Figures 2 and 3 show plots of the two types of laughter in terms of each acoustic representation for all data of the speaker JMA. From these plots we infer that type of laughter can be readily characterized by use of these acoustic features and will show the extent to which this can be achieved. Figure 2 shows first 8 parameters of JMA, and for example that 'p' (polite)

is characterized by relatively low maximum power, and that 'm' (mirthful) is characterized by relatively high maximum power. Most laughs are in the region of high maximum power and there is considerable spread of laugh categories across the whole of fmean-pmax dimensional feature space. Figure 3 shows last 8 parameters and note that 'p' (polite) is characterized by relatively high h1a3 value, which is a spectral tilt parameter representing differences in voice quality, and 'm' (mirthful) is characterized by relatively high duration and high No. calls. Since the data from speakerJMB and FAN show almost the same distribution as that of speaker JMA, their figures are omitted here (no individuals difference were found).

<**Figure 2 around here**>

<**Figure 3 around here**>

Principal Component Analysis (PCA) was used for analyzing and maximizing the combination of acoustic features across the speakers. The result from speakerJMB and FAN show almost the same as that of speaker JMA, and thus we report the PCA result for training data of the subject JMA. The proportion of variance from the first component to the fifth component (cumulative proportion of variance up to 70%) from a PCA rotation of these acoustic features shows that each component's contribution ratio is not indi-

vidually high, even for the first component, for all speakers. Table 4 shows the result of JMA's factor loadings. It reveals that the first principal component is largely related to fundamental frequency and No.call, the second to power, the third to spectral slope, and the fourth to F0moveAB.

<Table 4 around here>

4.5. Classification Trees

Classification Trees are a very useful tool for confirming finer details of contributing factors within the three parameters of fundamental frequency and power, min, max, and mean, that emerged from the principal component analysis.

We employed both Classification Trees and Support Vector Machines in our modeling; the former being relatively weak at classification but very useful for examining the contribution of the individual factors, and the latter being perhaps the strongest statistical classifier available for general use.

Figures 4 shows the results of growing and pruning a classification tree having 10 leaves for speaker JMA. Detailed formation of each tree differs according to speaker, but the important acoustic parameters are similar. These can be used to classify laughs according to a cascade of IF-THEN rules, giving total accuracy of 77% (JMA), 74% (JMB), and 90% (FAN) respectively.

Classification tree accuracies were measured for each test dataset. By observing the upper part of the tree, fmean, pmax, ppct, and dn (duration), the principal contributing features used to classify the two types of laughs can be determined.

<Figure 4 around here>

4.6. Both Speaker-dependent and Speaker-independent Classification by Support Vector Machine

Support Vector Machines are high-performance statistical classifiers. The SVM Type is C-classification, and kernel type is linear. Other system parameters are set to cost: 1, and gamma: 0.0625. The result of automatic discrimination using 15-fold closed (i.e., train and test on the same speaker) cross validation for JMA's mirthful (m) and polite (p) laughs we obtained 85% total accuracy. Training a Support Vector Machine on the same data gives a much more successful result, since it employs a total of 84 Support Vectors to predict the data, rather than the 10 terminal nodes determined by the Classification Trees. For JMB and FAN the same classification is implemented and total accuracy is 80% (JMB) and 92% (FAN). We split the data into training and test sets. The result of automatic discrimination on the test set for mirthful (m) and polite (p) laughs shows that we obtained

JMA; 75% (F-measure = 0.76), JMB; 87% (F-measure = 0.88), FAN; 84% (F-measure = 0.84) respectively.

In the speaker-independent classification, we trained with JMA (using training set) and tested with JMB and FAN (using test set). Two speaker's classification rates are JMB: 90% and FAN: 67% respectively. Good categorization was possible for JMB, however for FAN, classification rates is relatively low. It is probably caused by overfitting due to high dimensional acoustic parameters and thus we try to implement parameter reduction.

4.7. Parameter Reduction

Having a large number of predictor features usually results in better classification accuracy, but often at the cost of generalizability. Accordingly, we performed a Principal Component Analysis and used Classification trees to reduce the number of features used in the final model. As we mostly inspect the Table 4 and Figure 4, the important features were selected. the optimal combination of features was chosen from the first or second principal components, and from those featuring most commonly in the upper part of the Classification trees. We were able to confirm the usefulness of seven important acoustic features; fmean (or fmax), pmax, ppct, h1a3, duration, No.call, and F0moveAB. The other parameters were omitted from the set of acoustic

features used for the final training of the Support Vector Machines.

4.8. Classification by Support Vector Machine post Parameter Reduction

Following the above parameter reduction, we used a Support Vector Machine to predict the most likely category for each laugh token from its acoustics. The result of automatic discrimination using 15-fold cross validation for JMA's mirthful (m) and polite (p) laughs we obtained 86% total accuracy. For JMB and FAN the same classification is implemented and total accuracy is 86% (JMB) and 89% (FAN). It shows relatively high accuracies for each speaker compared to pre parameter reduction. The result of automatic discrimination using the test dataset for mirthful (m) and polite (p) laughs we obtained JMA; 79% (F-measure = 0.78), JMB; 89% (F-measure = 0.89), FAN; 79% (F-measure = 0.81) respectively.

4.9. Cross Prediction (speaker-independent) post Parameter Reduction

Table 5 shows the results of an open test across speakers, JMB, FAN, EMA, EFA, CMA, and CFA (mixing different native language and gender groups), training with JMA and testing with the others. All speaker's classification rates are over 70% (JMB: 85%, FAN: 74%, EMA: 93%, EFA: 79%, CMA: 86%, CFA: 86%). Good categorization was possible for each speaker

by using the seven acoustic features described above. However, difference in speaker-independent results before and after parameter reduction is not statistically significant. Therefore, we conclude that feature reduction could not actually help to significantly improve results.

We performed an error analysis for these SVM results restricted to polite tokens of two speakers, EMA (English male speaker) and CFA (Chinese female speaker) who represent difference of both gender, native languages, and age. A Student's t-test was conducted for each of the seven acoustic parameters. As a result, we found that pmax parameter differs between true (same test and training sample) polite laughter and error (false prediction or difference between test and training sample). According to this test, EMA's mean pmax "error" polite laughter: 54.99, "true" polite laughter": 71,56, p=7.11e-08 (signif) and CFA's mean pmax "error" polite laughter: 45.99, "true" polite laughter: 60.40, p=1.62e-05 (signif). This may be due to microphone impact noise since the power parameter was not normalized in the extraction process.

<Table 5 around here>

5. Discussion

This study evaluated classification of natural laughter for engagement sensing in natural speech data. In Experiment 1 we observed several types of laughs (mirthful, polite, derisive, and others) in a natural speech corpus, and two predominant types of laughter (polite vs. mirthful) were defined and categorized from a manual examination of the data and by perceptual labeling carried out by 20 Japanese subjects. In social communication (for Japanese at least, but probably more generally), people do not use hearty laughter with the same frequency that they utter polite laughs. We found that human laughter includes various laughs in conflict with stereotyped laughter [36], and we found many instances of the various types of laughter in our spontaneous Japanese speech.

This study reported an analysis of the acoustic features of these laughs. Global prosodic and laughter-specific acoustic features were extracted for the two types of laughter. These parameters were analyzed by Principal Component Analysis and Classification Trees to reduce the number of parameters. As a result of the analysis, we confirmed seven contributing acoustic features; mean value of fundamental frequency (fmean), maximum value of power (pmax), the position of the power maximum in relative percentage val-

ues (ppct), the difference between the first harmonic and the third formant (h1a3), duration of the laugh (dn), The number of calls in a bout (No call), and Pitch change between the first and the second call (F0moveAB). For both parameter plots and statistical analysis we found a difference between the two main types of laughter.

A Support Vector Machine was trained and tested using these seven features, and total classification accuracy was confirmed to be at least 85% with cross validation for speaker JMA. As a result of statistical analysis we reduced the number of parameters to seven dimensions. By observing the output of a principal component analysis and by use of classification trees some strong predictor parameters were chosen. After parameter reduction, open speaker tests across different discourse modes achieved approximately 70%.

We found certain individual differences and some strong similarities between people and tested both open and closed prediction methods. By reducing the number of parameters and using only the strongest and most general predictors we were able to obtain good results on cross-prediction tests for variety of speakers (cross-culture and personality). However, in case of EFA, her accuracy was low compared to other speakers in ESP corpus. She

seemed to be nervous during recording and thus she often laughs in a state of embarrassment that is difficult to classify into polite or mirthful laughs. Furthermore, speaker FAN data was recorded in various very different conditions as we mentioned in the introduction. That we can predict the type of laughter for her speech, when training on more constrained examples, indicates that this parameter reduction achieved high accuracy and allows high generalisation.

The present study justifies our belief that prosodic parameters are sufficiently and statistically different in the two types of laughter and that Machine learning can classify them efficiently. Scherer *et al.* [10] reported that the total accuracy of segmentation of laughter from natural discourse can be over 90%. This laughter detection is currently being integrated into a device to help people with autism spectrum disorders [31], who have difficulties understanding certain types of social functions. Finally we developed a Tcl/Tk based tool reflecting the result of these analyses. When the user speaks (in our present testing, usually acted laughs) into the microphone, and presses the analysis button, the system automatically displays the type of laughter (polite vs. mirthful) with an accompanying facial expression given by computer graphics. Support Vector Machine are used for classification pro-

cess. This tool might help train people with autism spectrum conditions to recognize human engagement in future. This is to be carried out as future work.

6. Acknowledgement

Most of this work was performed in the Applied Linguistics laboratory at Nara Institute of Science and Technology and later as joint work with the SFI FastNet Project at Trinity College Dublin.

References

[1] Petridis, S., Audiovisual Laughter Analysis, Ph.D. dissertation, University of London, 2011.

[2] Laskowski, K., and Burger, S., Analysis of the occurrence of laughter in meetings, In Proc. INTERSPEECH, pp. 1258—1261, 2007.

[3] Jung, W.E., The inner eye theory of laughter: Mindreader signals cooperator value, 2003.

[4] Vinciarellia, A., Pantic, M., and Bourland, H., Social signal processing: Survey of an emerging domain, *Image and Vision Computing 27*: 1743-1759, 2009.

[5] Ozonoff, S., and Miller, J. N., Teaching theory of mind: A new approach to social skills training for individuals with autism, *Journal of Autism and developmental Disorders 25*: 415-433, 1995.

[6] Golan, O., and Baron-Cohen, S., Systemizing empathy: Teaching adults with Asperger syndrome or high-functioning autism to recognize complex emotions using interactive multimedia, Develop. Psychopathology 18: 2006.

[7] Trouvain, J., and Schroder, M., How not to add laughter to synthetic speech, In Proc. The Workshop on Affective Dialogue Systems, Kloster Irsee, Germany, pp.229–232, 2004.

[8] Owren, M. J., Understanding Acoustics and function in spontaneous human laughter, Interdisciplinary Workshop on The Phonetics of Laughter, Saarbrcken, pp.4–5, August 2007.

[9] Truong, K. P., van and Leeuwen, D. A., Evaluating automatic laughter segmentation in meetings using acoustic and acoustic-phonetic features, Interdisciplinary Workshop on The Phonetics of Laughter, Saarbrcken, pp.49–53, August 2007.

[10] Scherer, S., Schwenker, F., Campbell, N. and Palm, G., Multimodal

Laughter Detection in Natural Discourses, *Transaction on Affective Computing*, pp.1–18, 2009.

[11] Kennedy, L. S., and Ellis, D.P.W., Laughter detection in meetings, NIST ICASSP 2004 Meeting Recognition Workshop, Montreal: pp. 118-121, 2004.

[12] Glenn, P. J, Current speaker initiation of two - party shared laughter, Research on Language & Social Interaction 25: 139-162, 1991.

[13] Jefferson, G., A technique for inviting laughter and its subsequent acceptance/declination, Everyday language: Studies in ethnomethodology 79: 96, 1979.

[14] Kangasharju, H., and Nikkot, T., Emotions in Organizations Joint Laughter in Workplace Meetings, *Journal of business communication 46*: 100-119, 2009.

[15] Shimizu, A., Yutaka, K., and Makoto, N., Hito wa naze waraunoka [In Japanese], *Kodan-sha*, 1994.

[16] Bachorowski, J. A., and Owren, M. J., Not all laughs are alike: Voiced

but not unvoiced laughter readily elicits positive affect, *Psychological Science*, 12(3): pp.252—-257, 2001.

[17] Hudenko, W. J., Stone, W., Bachorowski, J. A., Laughter differs in children with autism: An acoustic analysis of laughs produced by children with and without the disorder, *Journal of Autism and Developmental Disorders*, 39(10): pp.1392–1400, 2009.

[18] Laurence, D., and Laurence, V., Positive and negative emotional states behind the laughs in spontaneous spoken dialogs, Interdisciplinary Workshop on The Phonetics of Laughter, Saarbrcken, pp.37–40, August 2007.

[19] Bachorowski, J. A., Smoski, M. J., and Owren, M. J., The acoustic features of human laughter, *Acoustical Society of America*, pp.1581–1597, 2001.

[20] Sundaramb, S., and Narayananc, S., Automatic acoustic synthesis of human-like laughter, *Acoustical Society of America*, pp.527–535, 2007.

[21] Campbell, N., Kashioka, H., and Ohara, R., No laughing matter, In Proc. INTERSPEECH, pp.465–478, 2005.

[22] Carroll, J. M., and Russel, J. A., Do facial expressions signal specific emotions? Judging emotion from the face in context, *Journal of personality and social psychology 70*: 205, 1996.

[23] De Gelder, B., and Vroomen, J., The perception of emotions by ear and by eye, *Cognition & Emotion 14*: pp.289-311, 2000.

[24] The Expressive Speech Processing corpus: www.speech-data.jp

[25] Campbell, N., Whom we laugh with affects how we laugh. Proceedings of the Interdisciplinary Workshop on The Phonetics of Laughter: 61-65, 2007.

[26] Campbell, N., Differences in the speaking styles of a japanese male according to interlocutor; showing the effects of affect in conversational speech. *Differences 12*, 2007.

[27] Nishio, S., Koyama, K., and Nakamura, T., Temporal differences in eye and mouth movements classifying facial expressions of smiles, in Proc of Third IEEE International Conference on Automatic Face and Gesture Recognition: 206-211, 1998.

[28] Tanaka, H., Sakti, S., Neubig, G., Toda, T., Campbell, N., and Naka-

mura, S., Non-verbal cognitive skills and autistic conditions: an analysis and training tool, In Proc. 3rd IEEE CogInfoCom, pp. 41-46, Slovakia, Dec. 2012.

[29] Rietveld, T., Van Hout, R., Statistical techniques for the study of language behaviour, Berlijn: Mouton de Gruyter, 1993.

[30] Jokinen, K., Nishida, M., and Yamamoto, S., Eye-gaze experiments for conversation monitoring, In Proc. the 3rd International Universal Communication Symposium, pp. 303-308, 2009.

[31] Tanaka, H., Kashioka, H., and Campbell, N., Laughter as a gesture accompanying speech - towards the creation of a tool for the support of children on the autistic dimension, In Proc. GESPIN, Bielefeld, September 2011.

[32] Tcl/Tk Snack Toolkit www.speech.kth.se/snack/

[33] Hansen, H. M., Glottal characteristics of female speakers, Ph.D. dissertation, Harvard University, 1995.

[34] Tanaka, H., and Campbell, N., Acoustic features of four types of laughter

in natural conversational speech, In Proc. ICPHS XVII, pp. 1958-1961, August 2011.

[35] Ihaka, R., and Gentleman, R., R: a language for data analysis and graphics, J. Comp. Graph. Stat. 5: pp. 299-314. 1996. Available via http://www.R-project.org.

[36] Provine, R. R., and Yong, Y. L., Laughter: A stereotyped human vocalization, *Ethology* 89:115-124, 1991.

List of Tables

List of Figures

Figure 1: Log pitch contour and extracting method of two dynamic parameters F0moveAB and F0moveAN. These parameters need $F0avg2a$ which represents average logarithm of pitch within a first (A) call, and $F0tgt2b$ (Second (B) call) and $F0tgt2n$ (Final (N) call) which represents the pitch target mark at the end of each call by a simple regression coefficient. Pitch change between the first and the second call (F0moveAB) is calculated $F0avg2a - F0tgt2b$, and that between the first and the final call (F0moveAN) is calculated $F0avg2a - F0tgt2n$.

Figure 2: Showing first 8 parameters. JMA shows for example that 'p' (polite) is characterized by relatively low maximum power, and that 'm' (mirthful) is characterized by relatively high maximum power. Most laughs are in the region of high maximum power and there is considerable spread of laugh categories across the fmean-pmax dimensional feature space.

Figure 3: JMA shows a different distribution of categories across the different last 8 feature space which indicates that 'p' (polite) is characterized by relatively high h1a3 value, which is spectral tilt parameter correlated to voice quality, and 'm' (mirthful) is characterized by relatively high duration and high No. calls.

Figure 4: The Classification Tree for predicting laughs from JMA - with 10 leaves, using a different set of parameters and parameter ordering from that determined for JMA, starting from fmean, then taking into account dn (duration) and pmax (maximum power).

Tables

Table 1:

type	count	prop.	cumulative prop.
non-laughs	6999	none	none
polite	579	66%	66%
mirthful	244	28%	94%
derisive	49	5%	99%
others	4	1%	100%

Table 2:

	JMA	JMB	FAN	EMA	EFA	CMA	CFA
mirthful	129	127	196	5	44	57	5
polite	138	135	136	65	22	13	60

Table 3:

Features	Explanation
fmean	mean value of fundamental frequency
fmax	maximum value of fundamental frequency
fmin	minimum value of fundamental frequency
fpct	the position of the f0 maximum in relative percentage values
pmean	mean value of power
pmax	maximum value of power
pmin	minimum value of power
ppct	the position of the power maximum in relative percentage values
h1h2	the difference between the first harmonic and the second harmonic
h1a3	the difference between the first harmonic and the third formant
h1	the amplitude of first harmonic
a3	the amplitude of third formant
fvcd	the amount of voicing it contained
duration	duration of the laugh

Table 4:

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
fmean	0.406			-0.139	-0.360
fmax	0.426				-0.149
fmin	0.128	0.199	0.183	-0.249	-0.548
fpct	0.122			0.345	
pmean		-0.507			
pmax	0.223	-0.430			
pmin		-0.484			-0.180
ppct	-0.221		0.353	0.104	-0.184
h1h2	-0.258		-0.391		-0.158
h1a3	-0.312		-0.426		-0.154
h1	-0.287	-0.276	-0.312		-0.312
a3		-0.369	0.179	0.149	-0.184
fvcd		-0.202	0.167		0.494
dn	0.355		-0.411	0.113	
No.call	0.355		-0.382	0.106	0.148
F0moveAB				-0.620	
F0moveAN		-0.111		-0.580	0.177

Table 5:

JMB	mirthful	polite		EMA	mirthful	polite		CMA	mirthful	polite
mirthful	36	5		mirthful	4	1		mirthful	52	5
polite	6	24		polite	4	61		polite	5	8
FAN	mirthful	polite		EFA	mirthful	polite		CFA	mirthful	polite
mirthful	29	1		mirthful	43	1		mirthful	3	2
polite	15	16		polite	13	9		polite	7	53

Figures

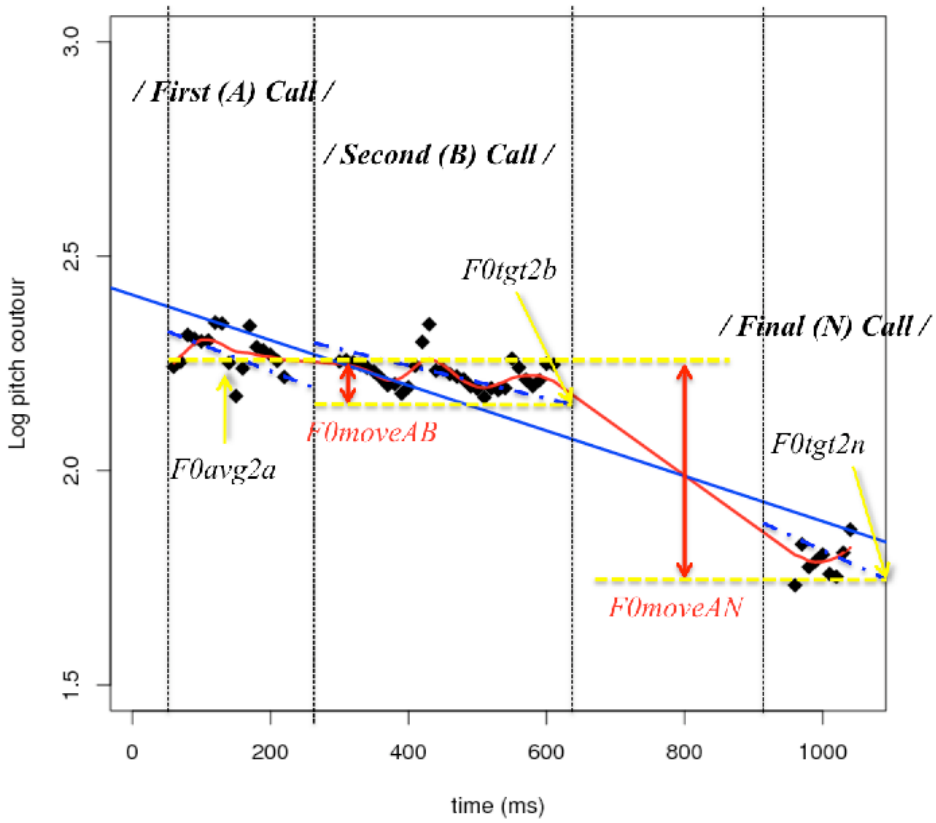

Figure 1:

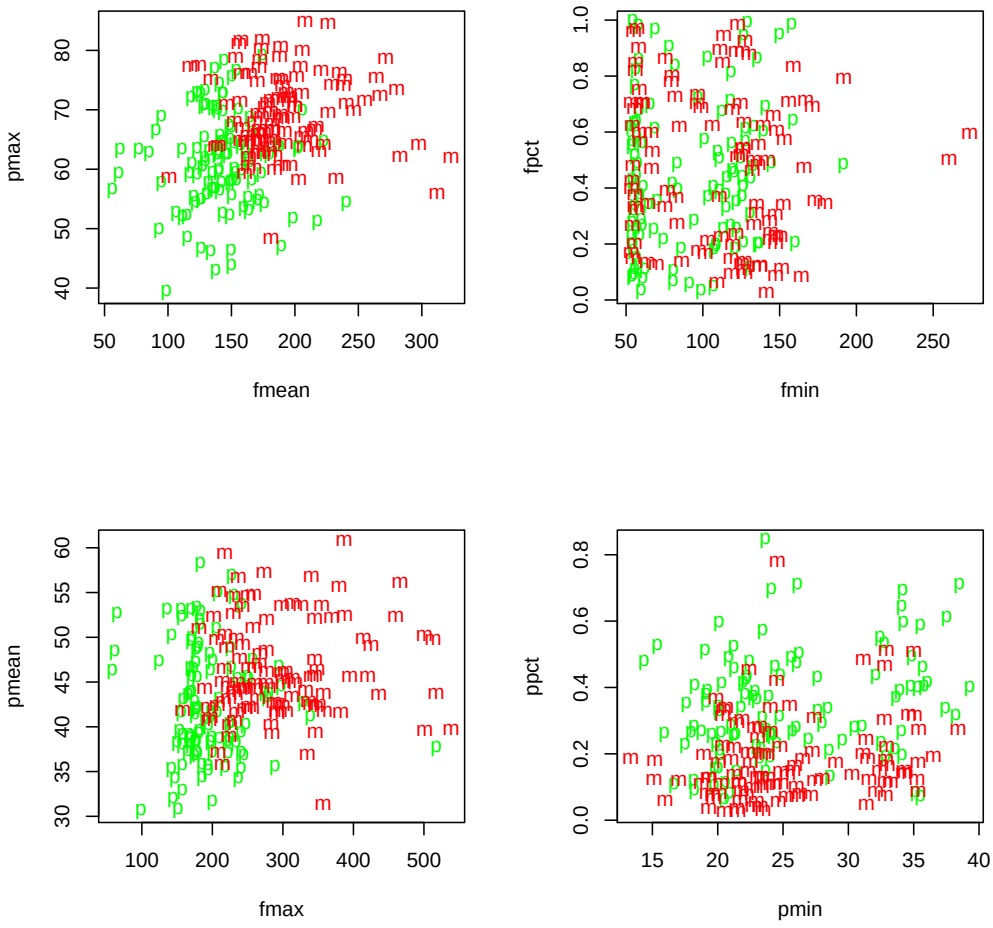

Figure 2:

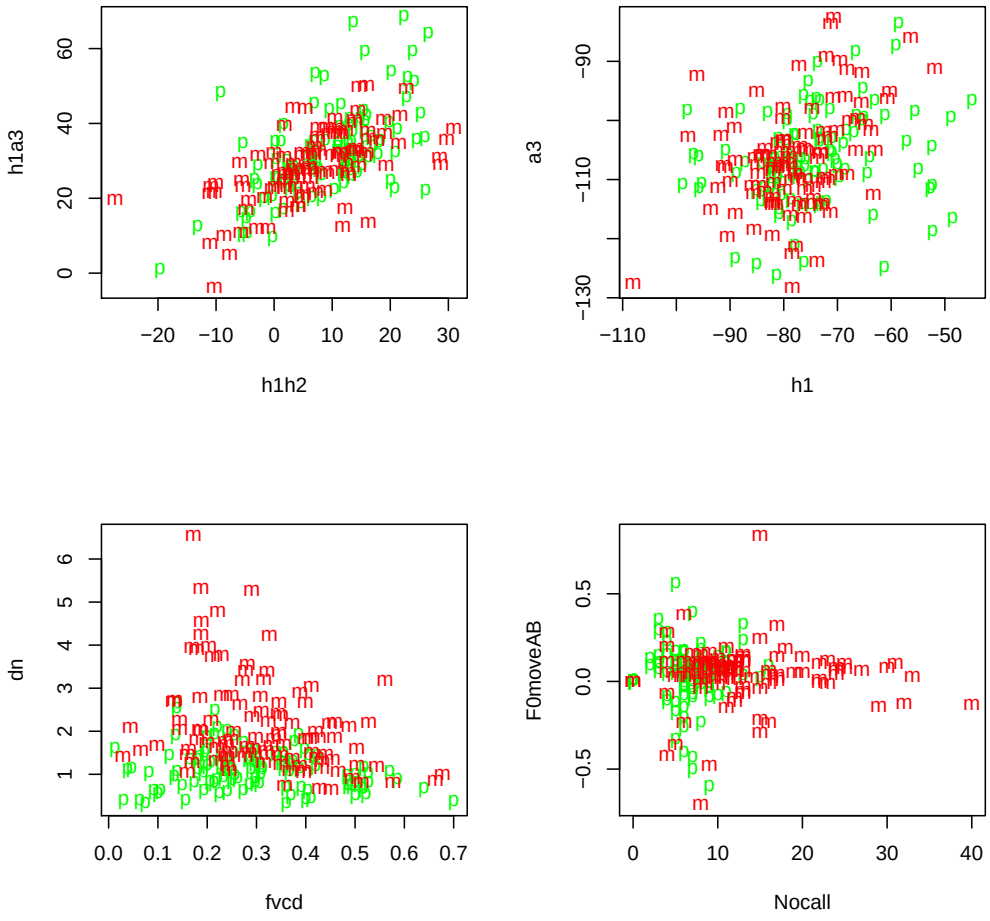

Figure 3:

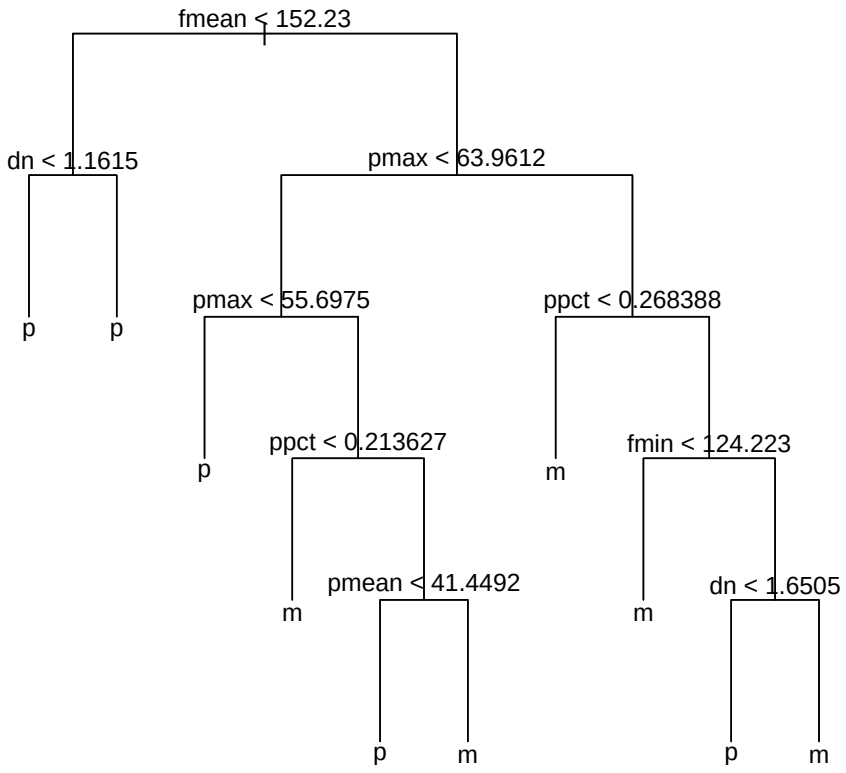
fmean < 152.23
dn < 1.1615
pmax < 63.9612
p
p
pmax < 55.6975
ppct < 0.268388
p
ppct < 0.213627
m
fmin < 124.223
m
pmean < 41.4492
m
dn < 1.6505
p
m
p
m

Figure 4: