Proceedings of the

Interdisciplinary Workshop on

The Phonetics of Laughter

Saarland University, Saarbrücken, Germany

4-5 August 2007

Edited by Jürgen Trouvain and Nick Campbell

PREFACE

Research investigating the production, acoustics and perception of laughter is very rare. This is striking because laughter occurs as an everyday and highly communicative phonetic activity in spontaneous discourse. This workshop aimed to bring researchers together from various disciplines to present their data, methods, findings, research questions, and ideas on the phonetics of laughter (and smiling). As a satellite event of the *16th International Congress of Phonetic Sciences* we think Saarbrücken was the appropriate place for this workshop.

We have been invited to act as guest editors of a special issue of the international inter-disciplinary journal *Phonetica* where selected contributions of submitted long versions shall be published next year. In addition, we hope that a mailing list open to the workshop participants and all other interested researchers will continue the exchange and stimulate further discussion on the phonetic aspects of laughter.

We were happy that Michael Owren (Georgia State University, Atlanta) agreed to give an invited talk on *Understanding Acoustics and Function in Spontaneous Human Laughter*. The planned tutorial plenary lecture on *The Phonetics of Laughter – A Linguistic Approach* could unfortunately not presented by Wallace Chafe (University of California, Santa Barbara) himself but Neal Norrick acted as presenter. All three keynote speakers deserve our thanks.

We would also like to thank the Fritz Thyssen Foundation for their financial support and the publisher S. Karger for their cooperation. Furthermore we are grateful to the Institute of Phonetics at Saarland University for providing the infrastructure and manpower for the workshop.

On this occasion we would like to thank very much all reviewers for their support. Special thanks go to Wallace Chafe for discussing various important issues during the planning of the workshop.

Saarbrücken and Kyoto, August 2006 Jürgen Trouvain and Nick Campbell

Reviewing Committee

- Kai Alter (Newcastle University Medical School, Newcastle-upon-Tyne)
- Jo-Anne Bachorowski (Vanderbilt University, Nashville)
- Nick Campbell (ATR, Kyoto)
- Hui-Chin Hsu (University of Georgia, Athens)
- Silke Kipper (Free University Berlin)
- Sabine Kowal (Technical University Berlin)
- Rod Martin (University of Western Ontario)
- Lucie Ménard (University of Québec, Montréal)
- Shrikanth S. Narayanan (University of Southern California, Los Angeles)
- Neal Norrick (Saarland University, Saarbrücken)
- Eva Nwokah (Communication Sciences and Disorders, University of North Carolina at Greensboro)
- Daniel O'Connell (Georgetown University, Washington, DC)
- Bernd Pompino-Marschall (Humboldt University Berlin)
- Béatrice Priego-Valverde (University of Aix-en-Provence)
- Willibald Ruch (University of Zürich)
- Marc Schröder (German Research Center for Artificial Intelligence, DFKI, Saarbrücken)
- Diana Szameitat (Newcastle University Medical School, Newcastle-upon-Tyne)
- Dietmar Todt (Free University Berlin)
- Jürgen Trouvain (Saarland University, Saarbrücken)

Table of Contributions

Invited Talk

UNDERSTANDING ACOUSTICS AND FUNCTION IN SPONTANEOUS HUMAN LAUGHTER

Michael Owren

Georgia State University, Dept. of Psychology
owren@gsu.edu

ABSTRACT

Laughter is often considered a stereotyped and distinctively human signal of positive emotion. Yet, acoustic analyses reveal a great deal of variability in laugh acoustics, and that changes in laughter sounds need not signal comparable changes in emotional state. There is simply not enough evidence to know whether laugh acoustics have specific, well-defined signaling value. However, there is evidence that laughter is deeply rooted in human biology. Great apes, for example, produce recognizably laugh-like vocalizations, and characteristic laughter sounds are produced by humans who are profoundly deaf. Based on acoustic form and likely phylogenetic history, laughter is argued to have evolved primarily as a vehicle of emotional conditioning. In this view, human laughter emerged because it helps foster and maintain positive, mutually beneficial relationships among individuals with genuine liking for one another. It is predicted to as easily have the opposite role among those who do not.

Series of similar vocal elements as a crucial acoustic structure in human laughter

Silke Kipper & Dietmar Todt

Institute of Biology: Animal Behaviour Group, Free University Berlin
silkip@zedat.fu-berlin.de, todt@zedat.fu-berlin.de

ABSTRACT

Among the many variable sounds in human laughter, vocalizations often contain series of vocal elements of similar acoustic properties. This study aims to elucidate whether such element series contain trajectories of changes in acoustic parameters that might be used to encode information, e.g. on the state of the signaller. We recorded bouts of laughter of adult humans (N = 17) and used a multi-parametric sound analysis to describe the acoustic parameters of vocal elements and their variation. We could show that these elements are distinguishable between individuals, but not necessary between female and male voices. We suggest that the series of similar elements with gradients in acoustic changes within laughter bouts might account for the stereotype and therefore predictable impression of laughter vocalizations.

Keywords: laughter vocalization, acoustic signaling, multi-parametric sound analysis.

ZUSAMMENFASSUNG

Innerhalb von Lachvokalisationen lassen sich Serien von Elementen mit ähnlichen akustischen Eigenschaften charakterisieren. Wir untersuchten, ob sich innerhalb solcher Elementfolgen Trajektorien akustischer Parameter-Änderungen beschreiben lassen, die zur Kodierung von Information genutzt werden können. Lachepisoden von 17 Erwachsenen wurden in einem multi-parametrischen Verfahren analysiert, um akustische Parameter von Elementen sowie deren Variabilität zu beschreiben. Die Elemente ließen sich anhand ihrer Eigenschaften den verschiedenen lachenden Personen zuordnen, das Geschlecht des Lachers war jedoch nicht in jedem Fall zu dekodieren. Wir schlagen vor, dass Serien ähnlicher Elemente mit geringen Veränderungen von Element zu Element sowie bestimmte Gradienten solcher Veränderungen den vorhersagbaren Höreindruck des Lachens hervorrufen.

1. INTRODUCTION

Some nonverbal human vocalizations such as infant crying or laughter contain series of acoustically similar elements [2,6,14]. It has been suggested that within such series, gradients of parameter changes might encode higher-order information that adds to the information encoded in the element structure [12]. Playback experiments using laughter recorded in natural settings or experimentally modified laughter did provide evidence that not only element characteristics, but also parameter changes within a series affect the evaluation of laughter by listeners [1,4,5,7,11].

The acoustic characteristics of human laughter have been investigated in several studies that either emphasized the stereotypy of the signal [8,9] or in contrary the enormous acoustic variability of laughter vocalizations [2,13]. Here, we investigate acoustic variation in laughter vocalizations by means of a multi-parametric sound analysis [10]. Comparing acoustic parameters of vocal elements for corresponding positions in a series will allow us to investigate the variability of laughter elements as well as rules underlying parameter changes within series of laughter elements.

2. METHODS

2.1. Recordings and acoustic analysis

We recorded spontaneous, unforced laughter during a reading task in which participants heard themselves with a short time delay (200 ms) through headphones (speech delayer: SPVZ, Geräte Zak, Simbach, Germany). This procedure led to problems in articulation which readily resulted in bursts of laughter (delay in playback was interrupted while recording laughter). 17 people volunteered to participate in the study (10 males, 7 females, mean age 36.6 ± 8.1 years). For recordings we used a DAT recorder (Sony TCD-

D10) connected to a Sennheiser ME 62 microphone.

Any laughter response that was comprised of a sequence of at least three elements was included in the analysis (overall, 178 bouts of laughter with 1515 elements). The acoustic properties of laughter vocalizations were analyzed using the sound analysis program Avisoft-SASLab Pro. R. Specht, Berlin (16-bit resolution, sampling rate 16 kHz). As a laughter *element*, we defined each discrete sound pattern that was not interrupted by pauses longer than 10 ms. Each *laughter bout* consisted of a number of elements within a wide range of acoustic characteristics and was finished either by a sound produced during inspiration or by the onset of speech. Laughter bouts often contained successions of similar elements. To operationally define these series, the following criteria was applied: *homotype element series* were all element successions where successive elements did not show more than 50 % difference in at least two of three acoustic parameters (duration, interval, and $F_{0\text{-max}}$, see Fig. 1 for illustration).

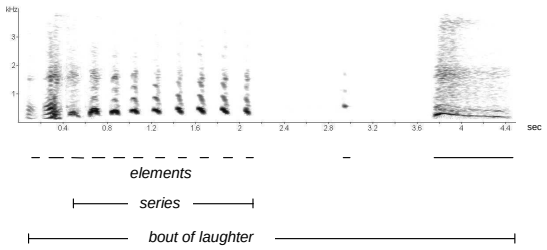


Figure 1: Spectrogram of a laughter vocalization (8 kHz, 16-bit). Acoustic structures are named below.

For each bout of laughter we measured the duration of the whole bout, of the series within the bout, and of each element, and for the latter, also the maximum value of the fundamental frequency. In addition, we obtained several measures for each element ('multi-parametric sound analysis') [10] by the following procedure: We digitized elements (8 kHz sample rate), and calculated spectrograms using a Fast Fourier Transformation (window size 512 sample points, overlap 93.75, time resolution 2 ms). Using the analysis program ConAn 0.9 (R. Mundry, Berlin), these spectrograms were used to calculate 115 measures on frequency and energy distribution over time and on the shape of the fundamental frequency for each element.

2.2. Data analysis and statistics

A discriminant function analysis (DFA) was applied to detect differences in laughter elements according to the two factors speaker identity and gender. In both analyses (speaker and gender), discriminant functions were calculated only on a subset of elements (internal elements, int.) whereas all elements were classified (external elements, ext., 'hold-out-sample method' [3]). Analyses were conducted with 7 acoustic parameters (statistic-based selection out of the 115 parameters measured), 14 subjects (7 male, 7 female), and 8 laughter series per subject. Loadings of parameters on the discriminant function were used to estimate their contribution to the discrimination of investigated classes.

All differences between groups or parameters were tested with non-parametric statistic tests.

To uncover trajectories of parameter changes within a laughter series, for some of the acoustic parameters measured, we calculated such changes according to the element's position within the series and introduced an additional measure (changes of 10 % or more) in order to roughly reflect perceptual abilities in humans. To assure comparability of results, this measure was only applied to series containing at least six elements.

3. RESULTS

3.1. Structure of laughter bouts

Laughter bouts were typically initiated by one or two 'singular' elements (i.e. non-repeated, with large variability in acoustic parameters). These were often followed by a succession of elements with predictable similarity, i.e. a homotype series. After this homotype series sometimes more singular elements followed and in the end often a sound produced during inspiration. Of all bouts analyzed, only 9 bouts (5%) produced by 7 different participants did not contain a homotype series. On the other hand, never did a bout of laughter contain more than one series.

In the majority of cases, homotype laughter series constituted the longest part of a laughter bout (Fig. 2). Same significant results were obtained by comparing the duration of temporal structures of the series for each participant, separately.

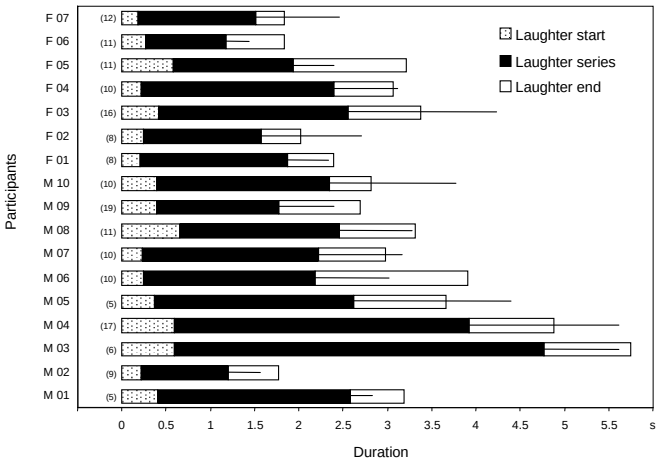


Figure 2: Temporal pattern of laughter bouts: series contribute most to a bout (Friedman one-way ANOVA for mean dur/subject, N=17, χ^2=34.0, p < 0.001).

Duration's of homotype laughter series did not differ between males and females, but we found considerable differences with respect to different participants.

3.2. Characterization of laughter elements in homotype series

The average rate of elements produced within series was 4.7 ± 0.8 /s. This measure did not differ between males and females, but between individuals, again. The results of the DFA also confirmed that individuality, but not gender, was distinguishable by the acoustic parameters of laughter elements (Table 1).

Table 1: Results of the discriminant function analyses and Binomial tests. Data were balanced by randomly selecting equal numbers (N=7) of subjects for genders. Within each analysis, subjects contributed equal numbers of internal elements.

	Subset	Correct classified (%)	Elements	*p*
Subject	int.	47.3	112	< .001
	ext.	19	807	< .001
Gender	int.	59	112	.01
	ext.	56	807	.07

Thus, laughter elements within series contained acoustic characteristics that, in their combination, made it possible to distinguish between laughing people. Parameters that were especially decisive for differences between subjects (i.e. such loading high on DF1) were those describing the frequency contour within an element: a measure of the fundamental frequency (FMedS), the slope of the element (as a measure of frequency contour, SlStEnd), the time until the minimal slope was reached (LocSlMin), and the maximal slope (SlMax).

3.3. Gradients of parameter changes within laughter series

For seven parameters (four that most effectively characterized individual differences and three that were efficient in eliciting different responses in playback experiments [4]), we correlated the measures of these parameters with the position of the elements in a series in order to describe gradients within acoustic parameters. We calculated Pearson's correlation coefficients for each parameter and each series, separately (all N=98, balanced for speaker and gender). Only the distribution of correlation coefficients for duration, interval, and fundamental frequency differed from a normal distribution, thereby pointing to gradients of parameter changes (Kolmogorov-Smirnoff-tests, duration: Z=1.47, p=0.03, interval: Z=2.60, p=0.001, fundamental frequency Z= 2.60, p=0.001, FMedS=1.37, p=0.05, SlStEnd, LocSlMin, SlMax: all Z<0.095, all p n.s.).
Correlation coefficients tended to be positive for intervals (22 positive out of 24 series with a significant correlation), whereas duration was clearly negatively correlated with element position (40 negative out of 46 series with a significant correlation). Element frequency within homotype series either decreased or increased (37 negative out of 48 significant correlation's). In other words: whereas intervals tended to get longer towards the end of a series, duration's declined in the course of a series. The fundamental frequency did show either decreasing or increasing gradients in different series.

The characterization of gradients by means of differences of at least 10 % between successive elements specified these results. Gradients showed a characteristic distribution over the course of the series (Fig. 3).

For the duration, increases did occur less often towards the end, whereas decreases tended to occur more often in the later transitions of a series. The interval was in most of the cases increasing within the course of a series. Pitch showed only few element transitions with differences of above 10 %. Increases of pitch occurred more often in the

beginning of series, whereas decreases occurred more often towards the end of a series.

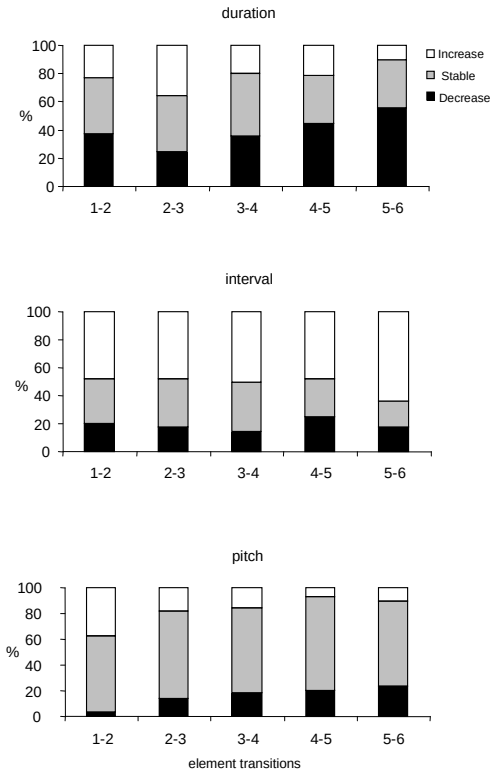


Figure 3: Proportion of gradients (increase, stable, decrease) in the element transitions of laughter by means of at-least-10-% differences between successive elements.

4. DISCUSSION

Inquiries into the structure of laughter elements and the dynamics of parameter changes in human laughter allowed us to show that this display includes sequences of similar vocal elements. Within such 'homotype series', acoustic parameters did only exhibit small changes from element to element. The comparison of acoustic features of successive elements uncovered some typical trajectories of parameter changes within laughter series. For example, there was a tendency for the duration of elements to decline within the series and for intervals to increase.

The range of parameter variability within the series of laughter elements is individually different, with measures characterizing the frequency shape of an element being especially decisive between subjects. Interestingly, neither these nor other investigated parameters showed systematic differences between female and male subjects. The failure to find such differences in our study might, on the one hand, be explained by methodological constraints. For example, whereas in other studies all laughter vocalizations performed during a whole bout or burst of laughter were analyzed, we considered only homotype laughter elements. On the other hand, listeners evaluating laughter sometimes did report difficulties to discriminate female and male voice characteristics (Kipper, unpublished data). Such a reduction of acoustic differences between male and female voices might point to the biological significance of laughter. There is a consensus that laughter is used as an 'in-group-signal' that serves to generate and maintain social bonds [e.g. 1,6,7]. A signal serving such a function might not be designed to accent differences between sub-groups.

We were able to extract specific rules of parameter variation within laughter series as given if parameter values shift from element to element within a sequence. Such rules, forming gradients or trajectories, have been documented in communicative systems of many animals and have been argued to serve communicative functions there [12].

5. CONCLUSION

In the present study we investigated the organization of laughter vocalizations applying the conceptual framework of homotype signal series. This allowed us to explain several features of human laughter and to extract rules of parameter variation. Studies on laughter vocalizations in different social settings are crucial to verify these results. At the same time such investigations will be the only way to raise our understanding on the signal design and variability of human laughter. These studies should include either side of the signaler/recipient system and consider the production and performance as well as the perception and evaluation of human laughter.

6. REFERENCES

[1] Bachorowski, J.-A., Owren, M.J. 2001. Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect. *Psychol. Science* 12, 252-257.

[2] Bachorowski, J.-A., Smoski, M.J., Owren, M.J. 2001. The acoustic features of human laughter. *J. Acoust. Soc. Am.* 110, 1581-1597

[3] Bortz, J. 1999. *Statistik für Sozialwissenschaftler*. Berlin: Springer.
[4] Kipper, S., Todt, D. 2003. Dynamic-acoustic variation causes differences in evaluations of laughter. *Percept. Motor Skills* 96, 799-809.
[5] Kipper, S., Todt, D. 2003. The role of rhythm and pitch in the evluation of human laughter. *J. Nonv. Behav.* 27, 255-272.
[6] Kipper, S., Todt, D. 2005. The sound of laughter – recent concepts and findings in research into laughter vocalisations. In: Garfitt, T., McMorran, E. & Taylor, J. (eds.), *The Anatomy of Laughter*. (Studies in Comparative Literature 8). Legenda: London, 24-33.
[7] Provine, R.R. 1992. Contagious laughter: laughter is a sufficient stimulus for laughs and smiles. *Bull. Psychon. Soc.* 30, 1-4.
[8] Provine, R. R., Yong, Y. L. (1991). Laughter: A stereotyped Human Vocalization. *Ethology, 89*, 115-124.
[9] Rothgänger, H., Hauser, G., Cappellini, A.C., Guidotti, A. 1998. Analysis of Laughter and Speech Sounds in Italian and German Students. *Naturwissenschaften* 85, 394-402.
[10] Schrader, L., Hammerschmidt, K. 1997. Computer-Aided Analysis of Acoustic Parameters in Animal Vocalizations: A Multi-Parametric Approach. *Bioacoustics* 7, 247-265.
[11] Sundaram, S., Narayanan, S. 2007. Automatic acoustic synthesis of human-like laughter. *J Acoust. Soc. Am.* 121, 527-535
[12] Todt, D. 1986. Hinweis-Charakter und Mittlerfunktion von Verhalten. *Z. Semiotik 8*, 183-232.
[13] Vettin, J., Todt, D. 2004. Laughter in conversation: Features of occurrence and acoustic structure. *J. Nonv. Behav.* 28, 93-115.
[14] Wermke, K., Mende, W., Manfredi, C., Bruscaglioni, P. 2002. Developmental aspects of infant's cry melody and formants. *Med. Engineering Physics* 24, 501-514.

FORMANT CHARACTERISTICS OF HUMAN LAUGHTER

D.P. Szameitat (1, 2), C.J. Darwin (3), A.J. Szameitat (2), D. Wildgruber (1), A. Sterr (2), S. Dietrich (1), K. Alter (4)

Universität Tübingen, Germany (1), University of Surrey, UK (2), University of Sussex, UK (3), University of Newcastle, UK (4)

d.szameitat@surrey.ac.uk, cjd@sussex.ac.uk, a.szameitat@surrey.ac.uk, dirk.wildgruber@med.uni-tuebingen.de, a.sterr@surrey.ac.uk, dietrich@cbs.mpg.de, kai.alter@newcastle.ac.uk

ABSTRACT

Although laughter is an important aspect of nonverbal vocalization, its acoustic properties are still not fully understood. Here we provide new data on the spectral properties of laughter. We measured fundamental frequency and formant frequencies of the vowels produced in laughter syllables. In accordance with theoretical predictions and prior observations laughter was mainly based on central vowels. Furthermore, laughter syllables showed higher formant frequencies than normal speech vowels; in particular F1 values could be as high as 1300 Hz for male speakers and 1500 Hz for female speakers. These exceptionally high F1 values might be based on the extreme positions adopted by the vocal tract during laughter in combination with physiological constraints accompanying production of a "pressed" voice.

Keywords: laughter, formant, vowel, nonverbal, F1.

1. INTRODUCTION

The acoustical signal of laughter has unique structural features. It consists of a series of repeated syllables produced on a staccato outward breath. Each syllable typically consists of a fricative (aspirated "h" sound) followed by a vowel element [25]. Moreover, laughter can be produced with extreme voice characteristics (e.g. squealing), with its pitch being up to 1245 Hz for male speakers and 2083 Hz for female speakers, respectively [1]. During production of such sound utterances the vocal tract can be under great physiological strain. Furthermore, during laughter the mouth can be opened very wide. This extreme articulation is likely to produce extreme acoustic consequences, such as very high F1 frequencies.

The most extensive study of the spectral properties of laughter was done by Bachorowski and colleagues [1]. However, although females should have higher formant frequencies than males because of their shorter vocal tract length [20], for some of the formants (i.e. F4 & F5) Bachorowski et al.'s outcomes [1] were not in line with this prediction. Since the authors themselves suggested that this result might be due to peculiarities of the analysis performed, there is a need for further analyses. Other studies that have investigated spectral properties of laughter examined either only a small number of subjects [3] or analysed only two formants [16].

Our study measured the fundamental frequency and the frequency of the first five formants of vowels in laughter syllables produced in various emotional contexts. We also determined vowel elements by comparing F1-F2 plots with Hillenbrand et al's speech vowel representation [11].

2. METHODS

2.1. Sound recordings

Eight professional actors (3 male/ 5 female) produced laughter in various emotional contexts (joy, tickle, *schadenfreude* [to laugh at another's misfortune], sneering). Recordings took place in a sound proof booth, using a DAT recorder (TASCAM DA-P) with a speaker-microphone (Sanyo MP-101) distance of circa 0.5 m. Recordings were digitized (16 bit / 48 kHz), normalized, and cut into individual laughter sequences.

2.2. Sound material

We excluded laughter sequences that contained words, interjections, or background noise, or were of short duration (< 3s) or low amplitude (with non-detectable pitch).

The stimulus set consisted of 125 laughter sequences (49 male) with 10-22 sequences per speaker. Formant frequency measurements were

obtained for 3932 laughter syllables (1689 male / 2243 female).

2.3. Acoustical analysis

Extraction of mean fundamental frequency (F0) and mean frequency of five formants (F1-F5) of each laughter syllable was conducted in Praat 4.02.04 [6]. Fundamental frequency analysis was based on an acoustic periodicity detection using an accurate autocorrelation method [5]. This method allows reliable pitch extraction also for vocalizations which are not fully voiced. Maximum pitch search range was determined by visual inspection, by overlaying the automatically extracted pitch contours with a narrowband FFT-based spectrogram (30 ms, Gaussian window, pre-emphasis +6 dB/octave). Formants were extracted performing a short-term spectral analysis (Gaussian-like window, LPC analysis, Burg algorithm, see [7, 21]), approximating the spectrum of each analysis frame by five formants. Ceiling of the formant search range was 5000 Hz for male and 5500 Hz for female speakers, respectively.

Laughter sequences were segmented in the time domain according to individual laughter syllables (burst of energy of (un)voiced exhaled breath having a single vocal peak). Boundaries of a syllable were determined visually in the amplitude-time spectrum (distinct rise of energy from background noise into a single vocal peak). For syllables with ambiguous outcome in the automatic formant extraction, formant-peak locations were examined by visual inspection on a random basis. For this, the automatically detected formant bands were overlaid with a broadband FFT-based spectrogram (5 ms, Gaussian window, pre-emphasis +6 dB/octave). Formant measurements were not taken from laughter syllables which were unvoiced, produced with closed mouth, or where spectral measurement extraction was uncertain.

To determine vowel quality of the laughter syllables, F1-F2 plots were calculated for each individual speaker and mapped with the speech vowel representation according to Hillenbrand et al. [11].

3. RESULTS

Table 1 shows average fundamental frequency and formant frequency measurements for laughter syllables produced by male (1689 syllables) and female (2243 syllables) speakers. Statistical tests revealed that in all six acoustical parameters female speakers had higher frequency values than male speakers (independent-samples t-tests, $t(6) = 2.657 – 5.027$, all $p<.05$, Bonferroni-corrected).

Table 1: Frequency measurements for fundamental frequency (F0) and first five formants (F1-F5) for male and female speakers. s.d. standard deviation.

[Hz]	Females	s.d.	Males	s.d.
F0	476	107	199	8
F1	924	128	728	11
F2	1699	93	1530	71
F3	2995	89	2700	58
F4	3842	152	3472	179
F5	4600	117	4184	264

Surprisingly, in 26% of all vowel elements in laughter syllables F1 frequencies were higher than 1000 Hz (n=1021), with male speakers showing maximal values up to 1300 Hz and female speakers up to 1500 Hz. Thus, first formants of several laughter syllables had exceptionally high values in comparison with speech vowels [e.g. 11, 20]. These syllables very often sounded as though they had been produced with a hard or "pressed" voice.

According to Hillenbrand et al.'s [11] standard-vowel-space-representation vowel elements of female speakers fall mainly into the (Λ) and (ɑ) range, with some vowel elements falling in the (ɪ), (e), (æ), (ɛ), (ɝ), (c) and (U) range (Fig. 1). Vowel elements of male speakers fall mainly into the (ɝ), (Λ) and (ɑ) range, with some vowel elements falling into the (i), (ɪ), (e), (æ), (ɛ), (c), (U) and (o) range (Fig. 2).

Analysis on the basis of individual male speakers revealed that all but c.10 of the vowel elements falling into the (i), (ɪ), (e) and (æ) range had been produced by the same male speaker. Thus, laughter syllables were predominantly based on central vowels, with vowel height varying from mid (ə) to open (a), probably because of changes in jaw opening.

Analysis of vowel quality according to speaker identity revealed that differences between vowel elements are based mainly on speaker identity. In other words, individual speakers tend to use a constant set of vowel elements (low intra-personal variability).

Figure 1: F1-F2 plot for female speakers with vowel representation according to Hillenbrand et al. [11].

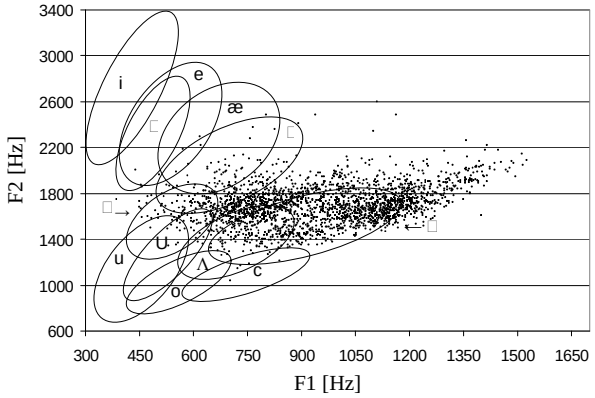
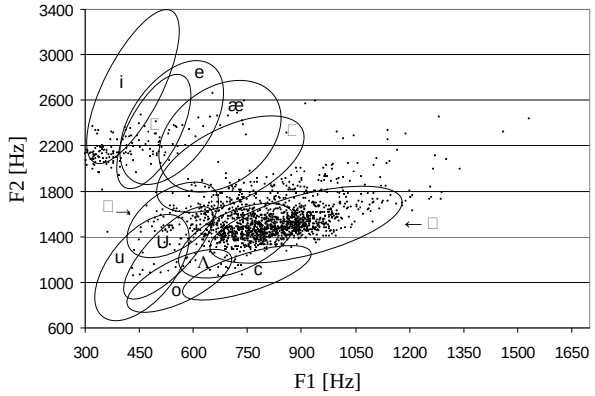


Figure 2: F1-F2 plot for male speakers with vowel representation according to Hillenbrand et al. [11].



4. DISCUSSION

4.1. Fundamental frequency

The mean F0 was 199 Hz for male and 476 Hz for female speakers, respectively, which is well within the range of previously reported F0 for laughter (mean-F0 range males (females) 160-502 Hz (126-424 Hz) [1, 3, 4, 14, 16, 17, 19, 22, 23]). Thus, our data are in accordance with the finding that fundamental frequency in laughter is higher than in speech [17-19, 23, 24].

4.2. Formant frequencies

While frequency measurements of the second to fifth formants fell within the range previously reported for laughter [1, 16] the first formant showed much higher frequencies than expected (here: 924 Hz (females), 728 Hz (males); Bachorowski et al. 2001: 653 Hz / 535 Hz).

High F1 values could be due to erroneous formant extraction. For instance, in high pitched sounds harmonics are widely spaced, so that the fundamental frequency can be higher than the actually articulated F1. The second formant may then be measured as F1. However, this artefact is unlikely to be the reason that we have obtained such high F1 values. Our high-F1 syllables were not particularly high pitched, but were characterised by a wide range of F0 values (for all F1 > 1000 Hz, females: range 81-1486 Hz, n= 878; males: range 155-404 Hz, n=50). In addition, visual inspection of broadband spectrograms of a random selection of very high-F1 syllables showed sufficient energy in lower frequency bands for an actually lower F1 to have been revealed. Finally, if the true F1 had been missed and F2 consequently identified as F1, then all following formants should be much higher as well (Paul Boersma, personal communication), but this was clearly not the case. Alternatively, so-called pseudo formants (reflecting turbulences in the air flow) may account for the high F1 values. However, pseudo formants are characterised by a high formant bandwidth (>500 Hz [13]), which we observed only in 3.5% of the high-F1 syllables. In addition, almost all examined syllables showed clear harmonic structure. Taken together, it seems very unlikely that the high F1 values are caused by erroneous analysis.

Another cause of the high F1 values may be found in physiological changes in the vocal tract. Firstly, lowering the jaw results in a raised F1 [26]. For instance, soprano singers can raise their F1 up to approx. 1050 Hz (to tune it to their F0) by opening the jaw very wide [12]. Secondly, certain voice qualities associated with narrowing the pharynx lead to a raised F1 [15]. Remarkably, most of the high F1 syllables were produced with a “pressed” voice which may well stem from physiological constraints in the pharyngeal region, such as a lower pharyngeal constriction. Therefore, it seems likely that the currently observed high F1 values are the result of a combination of wide jaw opening and pharyngeal constriction.

A possible explanation why other studies have not yet identified such high F1 frequencies for human laughter is that they may have used laughter which was less expressive, i.e. laughers’ arousal may have been lower than for our material. For instance, in the study of Bachorowski et al. [1] subjects laughed while watching funny video clips in an experimental setting, partly being together with strangers (for a similar approach see [16]). These circumstances may have inhibited the subjects' laughter response. This inhibition may have led to less extreme articulation, and conse-

quently a lower F1. In contrast, in our study actors were asked to put themselves fully into the emotional contexts given in the instructions, so that laughter might have been produced more expressively.

Another reason might be that in the current study the stimulus set was based on laughter produced by actors, and therefore might differ in its acoustical properties in comparison to spontaneously emitted laughter. Exhaustive acoustical analysis (not reported) revealed that the acoustical properties of the recorded laughter of our study showed no fundamental differences to recent findings for spontaneously emitted laughter [27] (for a similar finding see [2]). The only exception was the longer duration of the laughter, which was introduced by explicit instruction given to the actors since for further planned studies laughter sounds with longer duration were needed. Empirical tests, investigating if people can tell the difference between spontaneous emitted laughter and laughter produced by actors could give new insights on how representative the latter is of human laughter, in general.

A final explanation is that laughter was recorded in a variety of different emotional contexts, which leads to the fact that laughter was emitted with a variety of different voice characteristics [27].

4.3. Differences in speaker sex

Fundamental frequency was higher in female than in male speakers (cf. [22, 23]) with F0 being up to 1765 Hz for female speakers and 595 Hz for male speakers, respectively ([1]: males (females) 1245 Hz (2083 Hz); see also [18] for children: 3200 Hz).

For all five formants females had higher average frequencies than males, which is in accordance with females having a shorter vocal tract than males [20]. Therefore, the current data contradict some of the previously reported findings [1, 16] which found for some of the formants either no differences, or even higher frequencies for males than for females.

4.4. Vowels

Regarding the mapping of the F1-F2-plots for laughter with the speech vowel representation according to Hillenbrand et al. [11] it should be noted that both data sets consist of different speakers, hence different vocal tract lengths. Therefore, the direct comparison may be prone to some misidentification of vowels. However, outcomes of IPA transcription confirmed our results that mainly central vowels are produced in laughter.

The finding that our laughter consisted predominantly of central sounds is in line with the general hypothesis of Ruch and Ekman [25] and other recent data [1, 22]. The use of central vowels is in accordance with physiological constraints accompanying production of laughter: the vocal tract is in a relaxed position, moreover, raised lip corners and wide jaw opening leave little room for articulation [24]. However, some of our laughter syllables were non-central sounds, as also reflected in previous work [3, 8, 25]. The reason for the production of non-central vowels is not fully understood. Ruch suggested that non-central vowels may be indicators of different emotional qualities underlying the laughter [24]. However, recent findings [27] are not in line with this prediction for the emotional connotations of the laughter investigated in the present study. Furthermore, non-central vowels are also produced when people laugh in a single behavioural context [1], therefore variability in the emotional or situational context seems not to be the leading factor for the production of non-central vowels. Alternatively, use of non-central vowels might be related to intra-individual differences. Previously, it was speculated that each person has their own characteristic laughter sound [8, 10, 18]. This hypothesis is supported by our data, as we found that individual speakers tended to use a constant set of vowel elements, but inter-individual variability was high. To fully understand the use of non-central vowels further investigation is needed.

5. CONCLUSION

In conclusion, these findings indicate that (i) laughter syllables are predominantly formed with central vowels, although others can occur; (ii) formant frequencies show typical gender effects with higher frequencies in female speakers; (iii) compared to speech production, the first formant of laughter vowels is occasionally characterized by exceptionally high frequencies which may be the result of a wide jaw opening and/or pharyngeal changes in "pressed" voice; (iv) the vowel elements during laughter showed a relatively stable individual pattern, whereas the between subject variability was considerably higher.

6. ACKNOWLEDGEMENTS

We would like to thank the Marie Curie Foundation and the German Research Foundation (Deutsche Forschungsgemeinschaft) for financial support (D.P. Szameitat), DFG AL 357/1-2 (K. Alter) and DFG WI 2101/2 (D. Wildgruber).

7. REFERENCES

[1] Bachorowski, J.-A., M.J. Smoski, and M.J. Owren. 2001. The acoustic features of human laughter. *Journal of the Acoustical Society of America*, 110(3), 1581-1597.

[2] Bea, J.A. and P.C. Marijuán. 2003. The informal patterns of laughter. *Entropy*, 5, 205-213.

[3] Bickley, C. and S. Hunnicutt. 1992. Acoustic analysis of laughter. *Proc. Int. Conf. Spoken Language Process*, 2, 927-930.

[4] Boeke, J.D. 1891. Mikroskopische Phonogrammstudien. *Pflügers Arch. f. d. ges. Physiol.*, 50, 297-318.

[5] Boersma, P. 1993. *Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound*, in *IFA Proceedings of the Institute of Phonetic Sciences.* p. 97-110.

[6] Boersma, P. and D. Weenink. 2003. *Praat: Doing phonetics by computer (Version 4.02.04), http://www.praat.org*.

[7] Childers, D.G. 1978. *Modern spectrum analysis*. 1978, New York: IEEE Press.

[8] Edmonson, M.S. 1987. Notes on laughter. *Anthropological Linguistics*, 29(1), 23-34.

[9] Ekman, P. 1997 *What we have learned by measuring facial behavior*, in *What the face reveals*, P. Ekman and E.L. Rosenberg, Editors. Oxford University Press: New York. p. 469-485.

[10] Fry, W.F. and C. Hader. 1977. The respiratory components of mirthful laughter. *Journal of Biological Psychology*, 19, 39-50.

[11]. Hillenbrand, J., et al. 1995. Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97(5), 3099-3111.

[12] Joliveau, E., J. Smith, and J. Wolfe. 2004. Tuning of the vocal tract resonance by sopranos. *Nature*, 427, 116.

[13] Kienast, M. 2002. *Phonetische Veränderungen in emotionaler Sprechweise*. 2002, Aachen: Shaker Verlag.

[14] La Pointe, L.L., D.M. Mowrer, and J.L. Case. 1990. A comparative acoustic analysis of the laugh responses of 20- and 70-year-old males. *International Journal of Aging and Human Development*, 31(1), 1-9.

[15] Laukkanen, A., E. Björkner, and J. Sundberg. 2004. Throaty voice quality: subglottal pressure, voice source, and formant characteristics. *Journal of Voice*, 20(1), 25-37.

[16] Milford, P.A. 1980. *Perception of laughter and its acoustical properties*, in *Department of Speech Communication*. Pennsylvania State University: Pennsylvania.

[17] Mowrer, D.E., L.L. LaPointe, and J. Case. 1987. Analysis of five acoustic correlates of laughter. *Journal of Nonverbal Behavior*, 11(3), 191-199.

[18] Nwokah, E.E., et al. 1993. Vocal affect in three-year-olds: a quantitative acoustic analysis of child laughter. *Journal of the Acoustical Society of America*, 94(6), 3076-3090.

[19] Nwokah, E.E., et al. 1999. The integration of laughter and speech in vocal communication: a dynamic systems perspective. *Journal of Speech, Language, and Hearing Research*, 42, 880-894.

[20] Peterson, G.E. and H.L. Barney. 1952. Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, 24(2), 175-184.

[21] Press, W.H., et al. 1992. *Numerical Recipes in C: the art of scientific computing*. 2nd ed. 1992, New York: Cambridge University Press.

[22] Provine, R.R. and L.Y. Young. 1991. Laughter: a stereotyped human vocalization. *Ethology*, 89, 115-124.

[23] Rothgänger, H., et al. 1998. Analysis of laughter and speech sounds in Italian and German students. *Naturwissenschaften*, 85, 394-402.

[24] Ruch, W. 1993 *Exhilaration and humor*, in *Handbook of emotions*, M. Lewis and J.M. Haviland, Editors. Guilford Press: New York. p. 605-616.

[25] Ruch, W. and P. Ekman. 2001 *The expressive pattern of laughter*, in *Emotion, qualia, and consciousness*, A. Kaszniak, Editor. Word Scientific Publisher: Tokyo. p. 426-443.

[26] Sundberg, J. and J. Skoog. 1997. Dependence of jaw opening on pitch and vowel in singers. *Journal of Voice*, 11(3), 301-306.

[27] Szameitat, D.P. 2006. *Perzeption und akustische Eigenschaften von Emotionen in menschlichem Lachen. Unpublished PhD thesis.* University of Tübingen: Tübingen.

STATES OF THE LARYNX IN LAUGHTER

John H. Esling

Department of Linguistics, University of Victoria, Victoria, BC, Canada
esling@uvic.ca

ABSTRACT

Although laughter can occur in various sequential (horizontal) patterns that have been well described, the voice quality components that can occur in laughter also need to be described in order to constitute a vertical array of possible phonetic contrasts. These components have been referred to as 'states of the larynx' and interact with pitch and with the 'segmental' components of laughter in predictable ways. Changes in laryngeal state can influence the phonetic shape of the voiceless/voiced parameter. Alternations between sources of periodic vibration also play a role in the description of laughter, where these laryngeal components might not otherwise play a role in the phonetics of the given language. Canonical profiles of the principal states of the larynx during various episodes of laughter are demonstrated.

Keywords: states, glottis, larynx, phonation, laryngeal constriction.

1. INTRODUCTION

Recent research into 'voice quality settings of the larynx' and new taxonomic frameworks for interpreting the role of the laryngeal articulator in contributing to vocal quality make it imperative that the accurate phonetic description of 'states of the larynx' in laughter not be taken lightly. Traditionally referred to as 'states of the glottis,' the new paradigm outlines 13 cardinal postures that characterize the positioning of the articulatory structures of the glottis and of the supraglottic laryngeal constrictor mechanism [9, 5]. A reliable depiction of both laryngeal levels is essential because of the likelihood of rapid fluctuations in the control of airflow through the larynx during incidents of laughter and because the states that are adopted during laughter may not be the same as postures typical of an individual's normal speech. Laryngoscopic video movies of the appearance of each canonical state of the larynx during different types of production of laughter are investigated and presented in the form of video files. It is proposed that modifications and combinations of these basic states, along with respiratory airflow control and timing, are key parameters which together may be used to categorize different types of laughter. The basic list of 10 states of the larynx (excluding for the moment as phonatory targets the static, non-continuous states of prephonation, glottal stop, and epiglottal stop) adapted from Esling (2006a) [5] is:

- breath (abduction)
- modal voice (adduction/phonation)
- falsetto (adduction/phonation plus longitudinal stretching – high pitch)
- breathy voice (abduction plus adduction)
- whisper (abduction plus laryngeal constriction)
- whispery voice (abduction plus adduction plus laryngeal constriction)
- creaky voice (adduction/phonation plus laryngeal constriction – low pitch)
- harsh voice – low pitch (adduction/phonation plus laryngeal constriction plus aryepiglottic trilling – low pitch)
- harsh voice – mid pitch (adduction/phonation plus laryngeal constriction – mid pitch)
- harsh voice – high pitch (adduction/phonation plus laryngeal constriction plus longitudinal stretching – high pitch)

2. TAXONOMY AND METHOD

Rather than take a full set of possible voice quality types as the basis for laryngeal activity during laughter, it is more economical to consider the states of the larynx as the basic postures, since each one isolates glottal components (as glottal shapes) from the effects of the laryngeal constrictor (as supraglottic shapes) [7, 4, 3]. Another essential distinction is the product of integrating the supraglottic category and the specification of larynx height. The non-constricted postures are: Lowered larynx voice (low pitch) and Faucalized voice (Lowered larynx, high pitch). The constricted postures are: Pharyngealized voice (raised larynx, low pitch) and Raised larynx voice (high pitch). The tabular relationships presented above are based largely on part (c) of the table of voice quality settings taken from [6] and reproduced in Table 1.

2.1. Voice quality parameters

Table 1: The laryngeal portion of the Table of Voice Quality Settings from Esling (2006b) [6].

Descriptive phonetic labels for voice quality settings

a) Oral vocal tract settings (not included here for the purposes of laughter description)

b) Laryngeal constrictor settings plus larynx height

LARYNGEAL CONSTRICTOR:	***Constricted:***	***Non-constricted:***
	Pharyngealized voice (raised larynx, low pitch)	Lowered larynx voice (low pitch)
	Raised larynx voice (high pitch)	Faucalized voice (Lowered larynx, high pitch)

c) Phonation types (glottal settings plus laryngeal constrictor)

Non-constricted:	***Constricted:***		
	Whisperiness:	*Creakiness:*	*Harshness:*
Breath	Whisper	Creak	
	Whispery creak	Harsh creak	
Breathy voice			
	Whispery voice	Creaky voice	Harsh voice, low pitch
Modal Voice	Whispery creaky voice	Harsh creaky voice	
	Harsh whispery voice	Harsh whispery creaky voice	Harsh voice, mid pitch
	Whispery falsetto	Creaky falsetto	
Falsetto	Whispery creaky falsetto	Harsh creaky falsetto	Harsh falsetto
	Harsh whispery falsetto	Harsh whispery creaky falsetto	
			Harsh voice, high pitch (force increased)

2.2. Method of exploration

In each case, the posture during the production of laughter is explored directly from above by means of a rigid orally-inserted laryngoscope and also by means of a flexible nasally-inserted fibreoptic laryngoscope. The subject in the case of initial exploratory observations with the rigid oral scoping technique was the author. The oral posture adopted during testing of laryngeal parameters was a close variety of schwa [ə] during transoral observation (while the close front vowel [i] is normally used during transnasal observation, designed to advance the tongue and to clear the view over the tongue into the pharynx). The basic laughter sequence adopted for the purposes of exploratory observations (in order to test the relationships between the laryngeal production of a laugh and the range of states outlined in sections 1 and 2.1) was an 'imitated-laughter' sequence with a canonical voiceless/voiced alternation, which in its unmarked form would consist of a breath+voice pattern with the voiceless glottal fricative being followed by the target vowel, i.e. [həhəhəhəhə] or [hihihihihi]. Pitch declined over the performance of the syllable string, so that the principal variable would remain the alteration of the laryngeal state itself. That is, pitch was intended to be modified only as a result of the varying states of the larynx. An example of this sequence – a laugh in falsetto

mode (stretched vocal folds) – is shown in video 1. The methodology follows previous experimental procedures with languages in which either phonatory register or pharyngeal setting interact with pitch phonemically [4]. The parallel to laughter is one of scale and expectation. Not every language makes extensive use of the laryngeal articulator in its phonology, as a register feature, or for secondary articulations. Those that do have had to be defined in relation to the model in section 2.1. In order for us to be clear about whether laughter varies across this same range of vocal parameters, and whether speech and laughter in a given language use these parameters in the same or different ways, we must apply the full set of auditorily specified (and articulatorily documented) reference parameters, whether or not the given language possesses such features in its speech. The intent of this approach is to elaborate the set of tools available for laughter research, allowing new questions to be asked, such as whether languages that 'speak' in a particular way also 'laugh' in a particular way.

Cases of authentic laughter were drawn from videos produced nasendoscopically that were available in the large set of recordings in the University of Victoria Larynx Research Project. Occasionally during the course of the filming procedure, a subject will laugh during conversation with the researcher or the physician who are carrying out the experiment. Videos were scanned for such haecceitous instances of laughter, which are just as unreflectingly spontaneous as any other unscripted parts of the event. Although the laughter itself is authentically spontaneous, using this adventitious methodology of data collection has the drawback of being sociolinguistically random. The subjects in the data pool represent different ethnic and language backgrounds, and no attempt has been made to determine the reasons why they might laugh the way they do. It may very well be that certain aspects of laughter are generically fixed, while others are individually variable, but without a dedicated methodology to dissect such fine distinctions, the current methodology will be restricted to speculations on what are assumed to be generalized aspects of laughter.

2.3. Consequences

Trying to superimpose a segmental sequence on an underlying long-term posture of the vocal apparatus invokes the voice quality paradox [1, 10]; even a long-term, ostensibly permanent posture can only be quasi-permanent and has intermittent components, some of which are phonetically more indicative of the underlying posture than others. So, maintaining a given laryngeal state over a recurrent voiceless/voiced segmental sequence will sometimes invoke one salient state of the larynx and sometimes another. For instance, maintaining a state of breath throughout the sequence would be unmarked in the case of [h] but would necessitate a consequent change in the value of the vowel to voiceless, [hə̥hə̥hə̥hə̥hə̥]. On the other hand, a breathy-voiced background setting, would induce the consonants of the sequence to become voiced and the otherwise modal vowels to acquire a breathy component, [ɦə̤ɦə̤ɦə̤ɦə̤ɦə̤] (shown in video 2). The same relationship applies to the vowels of whisper, which become voiceless but with whisper stricture [13]; whereas whispery voice affects the quality of the consonants so that they become a whispery-voiced variant of [ḥ], and the vowels are also whispery voiced. A harsh-voice posture at low pitch, where aryepiglottic trilling is present (video 3) may preserve the voiceless/voiced distinction by reshaping the sequence into what is in effect a voiceless/voiced trill sequence [ʜʢ̥] – or [ʜʢ̥ɘ̝] or [ʜʢ̥ɘ̝̃] to show that the voiced trilling component co-occurs with the vowel (or [ʜʢ̥i̝] or [ʜʢ̥ĩ̝] when the voiced component shares the lingual and labial components of [i]). Although the voiceless/voiced alternation appears to be generically inherent to the nature of laughter [11, 12], the consonant may not always have to be a fricative or trill. Certain states of the larynx may override the breath component and replace it with glottal stop or with a stronger stop closure. Extreme laryngeal constriction in the case of harsh voice at high pitch (shown with the unmarked voiceless onset in video 4) is an example of this – a potential case where the voiceless stricture (consonantal) component of a given individual's style of laughter might become a stop rather than a fricative.

3. RESULTS AND DISCUSSION

3.1. The larynx modelled canonically

Here is a possible remapping of the segmental sequence under the influence of the canonically prescribed laryngeal settings. Despite the long-term overlay, the basic voiceless/voiced opposition is preserved. It is assumed, crucially in voice quality

theory, that whispered contexts, as opposed to breath, will evoke the constricted laryngeal setting [cf. 13]; hence the contrast between [h] for breath (unconstricted) and [ḥ] for whisper (constricted) in their respective sequences. The only phonetic factor distinguishing mid-pitch harsh voice from high-pitch harsh voice is pitch (antero-posterior stretching of the glottal vibratory mechanism); but this added component of tension may be enough in some cases to alter the onset consonant in a laughter sequence from a continuant to a plosive. This effect has yet to be tested.

A word about the development of this taxonomy may be useful. 'Breath' and 'voice' (as in 'modal voice') are ancient terms in phonetics that distinguish whether voicing is not present or present during a given sequence. The laugh in falsetto mode (video 1) is equivalent to modal voice in that [h] is voiceless and the vowel adducted for voicing, except that falsetto entails longitudinally stretched vocal folds. If a laugh begins with high pitch and finishes with low pitch, then the first part of the laugh would be more similar to falsetto and the latter part of the laugh would be more similar to modal voice or even to creaky voice. These presumptions depend quite a bit on the language spoken and its sociolinguistic preferences for how laughing should be done. In any culture, however, there is likely to be a range of variety in styles of laughter that indicate different paralinguistic meanings. The 'states of the larynx' approach is designed to account for the changes in laryngeal posture that this range of styles might cover.

Breath is canonically distinct from 'whisper' as a phonetic trait. The breathy-voiced glottal fricative is [ɦ], and the usual diacritic for breathiness is two dots beneath the symbol. Whisperiness is marked by one dot beneath the symbol in conventional voice quality notation, e.g [ɦ̤], although this usage is somewhat paradoxical in the case of the vowel symbol since whisper implies greater laryngeal constriction than breath, not less. Creakiness implies greater laryngeal constriction than for modal voice, inducing a lowering of pitch (fundamental frequency) through the shortening of the vocal folds. Both creakiness and harshness employ the laryngeal constrictor mechanism [3] and can be marked by the same diacritic – a subscript tilde to designate 'laryngealization' – though the fine meaning of how that laryngealization is accomplished articulatorily differs in the two cases. Aryepiglottic trilling is designated here with the voiceless and voiced symbols for 'epiglottal' (i.e. pharyngeal) trills [3]. Thus, when [ʜ] becomes voiced it transforms into [ʢ] which is also accompanied by glottal voicing, hence the use of the tie bar over both symbols and the tilde under the vowel, indicating in this case not only laryngealization but also active supplementary vibration of the supraglottic aryepiglottic folds. Harsh voice at mid pitch and harsh voice at high pitch do not differ in transcription (resembling the case of model voice and falsetto) because the principal factor distinguishing them is the addition of longitudinal stretching to create high pitch. The essential difference between model voice/ falsetto and harsh voice mid/high pitch is the engagement of the laryngeal constrictor for the latter pair.

The following list gives the segmental (CVCV) sequence that each state of the larynx implies, transcribed in narrow phonetic detail:

- breath (relative abduction) [hə̰hə̰hə̰hə̰]
- modal voice (adduction/phonation) [həhəhəhəhə]
- falsetto (adduction/phonation plus longitudinal stretching – high pitch) [həhəhəhəhə]
- breathy voice (abduction plus adduction) [ɦə̤ɦə̤ɦə̤ɦə̤]
- whisper (abduction plus laryngeal constriction) [ḥə̬ḥə̬ḥə̬ḥə̬]
- whispery voice (abduction plus adduction laryngeal constriction) [ɦ̰ə̬ɦ̰ə̬ɦ̰ə̬ɦ̰ə̬]
- creaky voice (adduction/phonation plus laryngeal constriction – low pitch) [hə̰hə̰hə̰hə̰]
- harsh voice (adduction/phonation plus laryngeal constriction plus aryepiglottic trilling – low pitch) [ʜ͡ʢ̰ʜ͡ʢ̰ʜ͡ʢ̰ʜ͡ʢ̰]
- harsh voice (adduction/phonation plus laryngeal constriction – mid pitch) [ḥə̰ḥə̰ḥə̰ḥə̰]
- harsh voice (adduction/phonation plus laryngeal constriction plus longitudinal stretching – high pitch) [ḥə̰ḥə̰ḥə̰ḥə̰] or [ʔə̰ʔə̰ʔə̰ʔə̰]

3.2. The larynx in spontaneous laughter

It is possible to observe spontaneous laughter using laryngoscopic methodology. Laughter can occur during an experimental session, and the laughter is as spontaneous as the speech that occurs under the same conditions. When using a nasally inserted endoscope with proper technique, subjects do not consider that their speech is abnormal, and neither do listeners, although there may eventually be a build-up of mucus which could become equivalent

to speaking while having a slight cold, but this is usually not noticeable.

In this very preliminary sampling of laryngeal behaviour during spontaneous laughter and speech-laugh episodes filmed laryngoscopically, only the basic modes of phonation are observed to occur: modal voice, breathy voice, and falsetto. It is estimated that the subjects were not departing widely from their usual speaking voice qualities when breaking into laughter. It is perhaps significant that the filming circumstances (and experimental conditions) were not wildly hilarious to begin with and that the laughter could most probably be classified as nervous and/or polite. It certainly did not cover a range of possible types of laughter that one might expect from a professional actor performing a series of laughs for comedic effect, for example. In order to explain all possible eventualities in the use of the larynx in generating the classic abduction-adduction alternation of a laughter episode, it might be possible to ask an actor, or persons with particularly interesting target laughter, to perform their laughter under laryngoscopic examination. This tactic, however, has not been adopted here and is perhaps not well advised. It is more reliable in phonetic research of this sort to establish cardinal reference points generated under phonetically controlled conditions with clear auditory targets and then to share the auditory targets and their articulatory correlates with other phonetic judges who can listen to laughter data and categorize the various shifts in laryngeal state. This is the approach that has been taken in our parallel research into the earliest speech sound production by infants [8, 2], where laryngoscopic intervention would not be ethical. Similarly, our laryngoscopic work with adults speaking various languages [4] serves as the basis for identifying how the larynx functions; then the knowledge of the auditory/articulatory correlates of possible sound production types is applied to the description of data sets of infant sound production [2]. Our recommendation in the study of laughter repertories is to follow this same practice of phonetically instructed listening.

In our films of spontaneous laughter episodes, six instances have been isolated from three subjects, all female and in their 20s in this case, and examined as an introduction to the laryngeal behaviour of laughs. It is clear that the basic pattern of abduction (breath, in all of the cases observed here) and adduction (glottal voicing) is being respected.

For the first subject, pitch appears to be within the normal range of her speech; and in three cases, her laughter represents her normal speaking voice quality (essentially modal voice). In one case, perhaps due to the context of the task, where high pitch was the experimental target, her laughter was higher-pitched, approaching falsetto as a canonical referent.

The second subject's speech-laugh is breathier than her normal voice. Articulatorily, her glottis is wider open in the abduction phase and for longer periods than in the modal-voice laughs of subject 1. Airflow is also presumably greater, but this has not been measured. The type of laughter in this case could be characterized as breathy-voiced laughter.

The third subject also produces a laugh that is slightly different from her normal voice, perhaps only by its being exaggerated, but this would be a matter for experimentation to resolve. Her laugh appears to have higher-volume air flow than her speech (not measured) with articulatorily wide abduction phases. The context of the experimental task involved the production of a strong glottal stop just before the laugh, although this would also be a case where experimentation is needed to resolve the influence of context on the quality of subsequent laughter. Another possible factor in the analysis of this laugh is the noticeably long voiceless expulsion of breath at the onset of the laugh. This significant initial expending of subglottal air could have an effect on the pressure variables during the voicing components of the laugh. From a pitch perspective, there appears to be a shift upwards from the subject's normal pitch range, at the same time as breath is being forcefully expelled. This creates the conditions for the voicing quality to be both higher-pitched and breathy at the same time, which would not be predicted based on the normal distribution of breathy voice (usually in the lower end of the pitch range) [10, 3].

This final, sixth instance of unscripted spontaneous laughter from our preliminary laryngoscopic observations causes us to consider a number of points. First, it appears to be likely that a person's laughter can be produced in a different mode from their normal speaking voice quality. This is based on woefully limited data, from two subjects out of three. More significantly though,

the quality of subject 3's laugh suggests that airflow parameters will be critical in determining the kind of voicing that is produced during the adduction phase. Pitch is also recognized to be a powerful variable in altering the perception of phonetic quality; for example, the two constricted settings in Table 1(b) involve the same articulatory posture with only a difference in pitch level, and the two unconstricted settings in Table 1(b) involve the same articulatory posture with only a difference in pitch level. This last example in our set could not be called whispery, because the voiceless phases are too wide and therefore would have to be labelled breathy; and it is clearly not constricted either. Based on auditory phonetic analysis [10, 6], this laugh can be characterized as an example of faucalized voice (lowered larynx, high pitch), which is the opposite in posture to a constricted state. In faucalized voice, larynx position is low (as for lowered larynx voice) but pitch is high. In the video data, the supraglottic space is open, not constricted, the larynx can be seen to lower, and the antero-posterior distance remains relatively long, confirming (or at least visually illustrating) the auditory analysis.

In conclusion, it is perhaps worth emphasizing that states of the larynx are a critical component of the analysis of laughter since laughter inherently comprises, by definition, a rapid alternation between two distinct states of the larynx/glottis. This means that laughter is a phenomenon that is already identified on the basis of a contrast in laryngeal states. What those states are can also differ in voice quality, just as various speaking styles can differ in voice quality. So two distinct states, such as breath and voice, can also be influenced by an overlay of a supplementary quality that alters one or both of them. A change in the voice quality of laughter has implications for the segmental identity of its composite states, which can be retranscribed following phonetic principles. Another phonetic question to address is whether the aerodynamic components of the contrasting states in laughter are more exaggerated than in speech and therefore require redefinition from the norms identified for speech. Further linguistic questions can be asked once the superordinate voice quality and dependent segmental alternations have been identified, such as how laughter differs from non-laughter modes of an individual's speaking voice, how socially and regionally contrasting groups differ in styles of laughter, and how the acquisition of laughter is related to the acquisition of the speaking modality.

4. REFERENCES

[1] Abercrombie, D. 1967. *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

[2] Benner, A., Grenon, I., Esling, J.H. 2007. Acquisition of voice quality parameters: a cross-linguistic study. Paper presented at the 16th ICPhS, Saarbrücken.

[3] Catford, J.C., Esling, J.H. 2006. Articulatory phonetics. In: Brown, K. (ed), *Encyclopedia of Language and Linguistics* (2nd edn.) vol. 9. Oxford: Elsevier, 425–442.

[4] Edmondson, J.A., Esling, J.H. 2006. The valves of the throat and their functioning in tone, vocal register, and stress: laryngoscopic case studies. *Phonology* 23, 157–191.

[5] Esling, J.H. 2006a. States of the glottis. In: Brown, K. (ed), *Encyclopedia of Language and Linguistics* (2nd edn.) vol. 12. Oxford: Elsevier, 129–132.

[6] Esling, J.H. 2006b. Voice quality. In: Brown, K. (ed), *Encyclopedia of Language and Linguistics* (2nd edn.) vol. 13. Oxford: Elsevier, 470–474.

[7] Esling, J.H. 2005. There are no back vowels: the laryngeal articulator model. *Canadian Journal of Linguistics* 50, 13–44.

[8] Esling, J.H., Benner, A., Bettany, L., Zeroual, C. 2004. Le contrôle articulatoire phonétique dans le prébabillage. In: Bel, B., Marlien, I. (eds.), *Actes des XXVes Journées d'Étude sur la Parole*. Fès: AFCP, 205–208.

[9] Esling, J.H., Harris, J.G. 2005. States of the glottis: an articulatory phonetic model based on laryngoscopic observations. In: Hardcastle, W.J., Beck, J.M. (eds), *A Figure of Speech: A Festschrift for John Laver.* Mahwah, NJ: Lawrence Erlbaum Associates, 347–383.

[10] Laver, J. 1980. *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press.

[11] Trouvain, J. 2001. Phonetic aspects of 'speech-laughs'. In: Cavé, C., Guaïtella, I., Santi, S. (eds), *Oralité et Gestualité.* Paris: l'Harmattan, 634–639.

[12] Trouvain, J. 2003. Segmenting phonetic units in laughter. In: Solé, M.J., Recasens, D., Romero, J. (eds), *Proceedings of the 15th ICPhS.* Barcelona, 2793–2796.

[13] Zeroual, C., Esling, J.H., Crevier-Buchman, L. 2005. Physiological study of whispered speech in Moroccan Arabic. In: *Interspeech 2005 Proceedings*. Lisbon, 1069–1072.

‘SPEECH-SMILE’, ‘SPEECH-LAUGH’, ‘LAUGHTER’ AND THEIR SEQUENCING IN DIALOGIC INTERACTION

Klaus J. Kohler

Institute of Phonetics and Digital Speech Processing (IPDS), University of Kiel, Kiel, Germany
kjk AT ipds.uni-kiel.de

ABSTRACT

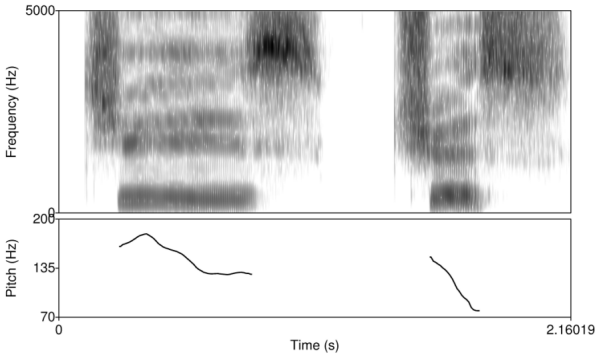
Laughing is examined auditorily and acoustico-graphically, on the basis of exemplary speech data from spontaneous German dialogues, as pulmonic air stream modulation for communicative functions, paying attention to fine phonetic detail.

1. INTRODUCTION

Phonetic analyses of laughing are relatively rare. Existing classification [2,11,12] differentiates *free laughter* and speech-synchronous forms that may be either *speech-laugh* [6], which is characterized by typical sequential laughter attributes coocurring with speech articulation, or *speech-smile*, where f0 is raised and long-term articulatory prosodies of, e.g., lip-spreading and palatalization are superimposed on speech events. *Speech-smile* is likely to be preceded and followed by the facial expression of a smile without sound, speech or paralinguistic vocalization, and is then only visible. This paper focuses on audio aspects of smiling only. For an analysis of smiling in the wider sense it would be necessary to have video recordings, which are not part of the database under discussion.

There has been controversial discussion as to whether speech-laugh and speech-smile form a continuum [12]. Sequences of speech-smile – speech-laugh (– free laughter) and the reverse order are attested in spontaneous speech corpora. But irrespective of this syntagmatic binding, these types of laughing phenomena are different production categories, and therefore they do not vary along one articulatory scale; they are also perceived as belonging to different categories and have different communicative functions. Free laughter and speech-laugh appear to be the expression of *amusement* and *high-key hilarity*, no matter whether real, acted, faked or ironical, although the manifestations may all be different because they can be recognised as such. Speech-smile, on the other hand, is more likely to be a signal of *happiness* and *low-key joy*. It may lead or trail the former, or it may stand on its own, for example as an expression of friendliness. A customer-friendly greeting *“tschüss”* (“bye”) at a shop cash desk may be spoken with a smile as in fig.1 and *audio_file_1.wav*, strengthening high frequencies in spectrum and f0.

Figure 1: *“tschüss”* [tʃyːs] smiling (left), [tʃʏs] neutral (right): spectrogram and f0 trace.



In the literature [2,10,11], the following parameters are listed for the phonetic categorization of different types of laughter:

- voicing, open mouth
- voicing, closed mouth, nasal exit
- voiceless nasal-cavity turbulence
- voiceless laryngeal and oral cavities turbulence
- vowel resonance: close – open, front – back
- pitch of voiced laugh bursts
- number of laugh bursts in a “bout” of laughter
- durations of laugh bursts and burst intervals.
- initial and final inhalations and exhalations are not always included in the analysis of bouts although they are important in the control of breathing for laughing.

Speaking has been described as modified (pulmonic exhalatory) breathing [1], involving complex supraglottal articulation as well as special subglottal pressure settings [5], peripherally supplemented by inhalatory pulmonic and other air stream meachanisms [9]. Pike [9] defined laughter as spasmodic articulation, produced by sudden movements beyond the control of the individual, in the same set as cough, sneeze, hiccough, or belching. This characterization is hardly adequate be-

cause laughter is controlled according to the speech function it is to subserve, even if it may be impressionistically described as spasmodic.

Laughing should be seen as another way of modifying breathing, largely exhalation, but involving inhalation as well, more so than in speaking. Research into the phonetics and communicative functions of laughing needs to analyse breath control (air stream direction, energy time course) as the basic phonetic element, which is modified in fairly simple ways glottally (vibrating/open/closed) and supraglottally (oral/nasal cavity, roughly positioned pharyngeal/oral/labial constrictions).

This is the reversal of the fine-grained phonation and supraglottal articulation superimposed on the relatively simple subglottal setting for speech, which is only modified under special circumstances such as a 'force accent' [3]. In other words, laughing should be analysed as modified breathing in its own right in parallel to speaking, rather than trying to apply speech categories, such as the distinctive vowels and consonants, or CV syllables, of a particular language, while neglecting the fundamental air stream control. Such an independent analysis of laughing will show up the correspondences and divergencies between the two ways of controlling the pulmonic air stream, and will then allow us to make insightful inferences as to how the two are combinable in speech-laughs.

Vocal tract resonances are no doubt important in colouring laughter in various ways for functional purposes, and will therefore need acoustic investigation, but these vocalic qualities are different, phonetically and functionally, from the vowel distinctions in the phonological system of the language. The latter are more numerous and more complex in distinctive feature composition, and serve the differentation of words. The resonances in laughter do not coincide with these qualities in phonological vowel systems and are more elementary, such as 'highish in the front region' vs. 'lowish or rounded in the back region' vs. 'central up-down and front-back' Their function is semantic and pragmatic to distinguish, e.g., 'giggle' and 'chuckle'. This is the same type of difference as vocalic qualities of hesitations, which do not coincide with phonetic ranges in phonological vowel systems either [7]. It is vocal tract resonance as a suprasegmental carrier as against a local segmental differentiator.

Of course, spontaneous laughter cannot be investigated in physiological laboratory experiments, so the direct observation of subglottal activity in laughing is precluded. Thus we have to deduce production, to a certain extent, from the acoustic result, and to this end, need to analyse the acoustic signal, coupled with acute auditory observation, in fine phonetic detail. Up to now, phonetic analysis of laughter has relied on rough acoustic measures, has hardly included the acoustic consequences of air stream direction, dynamics and timing, and has applied descriptive and inferential statistics too quickly to roughly analysed and tagged corpora on the basis of acoustic properties without due consideration of communicative function. The question of synthesizing laughter to make synthetic speech more natural-sounding is totally premature. What we need are studies of fine phonetic detail in dialogic interaction on the basis of exemplary spontaneous speech data. This paper provides a few results of such a (quite limited) investigation. Its aim is programmatic rather than a comprehensive descriptive account of a large database.

2. DATABASE AND METHOD

Two data sources have been used.

- A stereo recording of a dialogue session, consisting of 6 sub-dialogues, between two female speakers (institute secretaries), recorded in the Appointment-making Scenario with overlap [4], labelled but so far not published: **f06**.
- A stereo recording of two male speakers from the Video Task Scenario LINDENSTRASSE [4,8], **l06**, talking about differences in video clips presented to them separately.

In both cases, the speakers knew each other well, and they showed a high degree of spontaneity and naturalness.

In **f061**, speakers jm and mg have to arrange two 2-day business meetings in a 2-month period but cannot find mutually suitable dates because the experimenter inadvertently filled their respective calendars in such at way that there are not enough successive daily slots for both. The only solution mg can suggest is to have the two meetings immediately following each other, turning the two 2-day into one 4-day meeting. jm considers it a possibility but not an appropriate one. She finds this clash between the non-solvable task in the appointment-making game and the hardly conducive adjustment amusing, which she expresses by speech-laugh followed by subdued laughing. It is commented on by mg, with speech-laugh and laughter, as not being important in this kind of appoinment-making. mg's

amusement is in turn picked up by jm, leading to several laugh exchanges between the two speakers.

At the beginning of **l06**, speaker mpi sets the theme of "the utter stupidity" of the German TV soap series LINDENSTRASSE, capping his appraisal with hilarious laughter. The whole dialogue then revolves round exchanges on this theme between speakers mpi and tra about the episodes they have been presented with separately. This leads to several exchanges of laughing.

The recordings were listened to, and the laugh sections excerpted, in *cool edit*. The excerpted sections were then acoustically processed (spectrogram, f0, energy) in *praat* and descriptively analysed by close auditory and visual inspection.

3. RESULTS

3.1. Sequencing of 'speech-smile', 'speech-laugh', and 'laughter'

Any sequencing of the three laughing phenomena is possible. In an elaborate form, a speech-smile can develop into a speech-laugh and in turn into laughter, or contrariwise laughter continues into speech and then trails off as a speech-smile.

Figure 2: f061_jm *"glücklich"* + subdued laughter spectrogram, f0 (plain), energy (dotted).

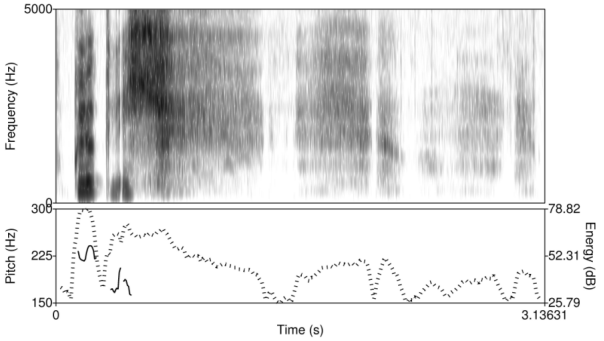


Figure 2 and *audio_file_2.wav* provide an example of the former order by speaker jm. Here *"glücklich"* ("suitable") is preceded by *"nicht so"* ("not so") and a breath intake, and shows a strong energy increase in the accented vowel as well as an f0 rise, indicating low-key joy over the clash between task and executability. This is followed by a renewed strong energy increase in the second syllable *"-lich"*, accompanied by a phonation (and f0) tremor, signalling incoming laughter. There is then long voiceless oral exhalation with high front vowel colouring before two oral cycles of voiceless long breathing in and short breathing out, energy decreasing progressively. This indicates subdued laughter, terminating the laugh section before resuming normal articulation for *"gut. also machen wir das so."* ("good. let's do it that way.").

Figure 3a: f061_mg *"ja"* + laughter + *"ja, gut"* + laughter: spectrogram, f0 (plain), energy (dotted).

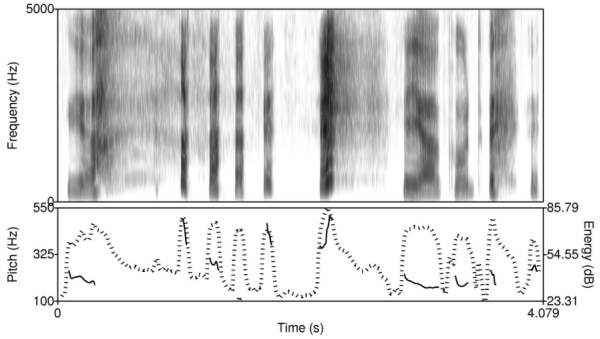


Figure 3b: f061_mg *"ja, gut"* vs. f065_mg*"na gut"* spectrogram and f0 trace.

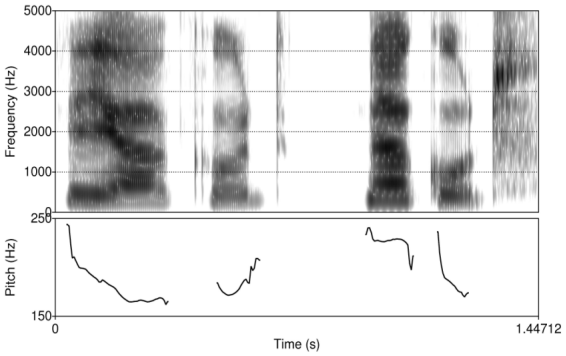


Figure 3a and *audio_file_3a.wav* provide an example of a more complex sequencing by speaker mg. *"ja."* ("yes") shows energy modulation: the energy rises into the vowel, then falls, and rises again, dividing the vowel into two. This is followed by strong voiceless exhalation, with open vowel resonance, setting in at the energy level of the sonorant vowel and trailing off over 400 ms. There are then, embedded in continuing exhalation, 4 voiced energy bursts of approximately the same duration, 70-80 ms, and of the same abrupt rise-fall of energy, evenly spaced, 140-170 ms, creating a very rhythmical pattern of strong laughter. The first 3 bursts have half-open central vowel resonance and descending f0, the 4th has high-front vowel resonance and a high upward f0 jump. The sequence is followed by a 400 ms pause and another, longer voiced energy burst, 120 ms, on an ingressive air stream, with an abrupt rise to a level 14 dB above the previous 4 bursts, accompanied by an abrupt f0 rise from 355 Hz to 512 Hz. The vowel resonance is less high and less front than the preceding burst.

This terminates the laughter and is followed by voiceless exhalation, which turns into speech with

a smile: *"ja, gut."* ("well, all right") has fronting throughout, most obvious in [u:] and [t]. The speech-smile then develops into a concluding short laugh with which the speaker hands over her turn: it consists of a short voice burst, trailing off in strong voiceless exhalation, followed by another much weaker voice burst, all of mid-central vowel resonance. This concluding laugh lacks the rhythmicity of a laughter pattern.

Figure 3b and *audio_file_3b.wav* compare mg's phrases *ja, gut,* of Figure 3a, and *na gut*, with and without a speech-smile, respectively. In the speech-smile, F1 and F2 of [u:] are raised, [t] has a higher locus frequency as well as an energy concentration of the burst at this raised locus, and f0 rises, thus high frequencies are strengthened.

Similar variation in the sequencing of the three types of laugh phenomena is found in the dialogue **l06** of the two male speakers (cf. 3.5 for examples and further discussion).

3.2. Air stream direction

As illustrated in 3.1, voiced and voiceless breathing occurs both egressively and ingressively in laugh turns. This is also found in **l06.** Final exhalation or inhalation should thus be treated as part of such turns and not be ignored in the analysis.

3.3. Oral and nasal air streams

Figure 4 and *audio_file_4.wav* illustrate nasal and oral air streams in laughs of f061_jm. After the utterance *"ja, dann hätten wir's, ne."* ("well, that's it, isn't it"), there is oral + nasal exhalation, which is followed by an oral closure. In turn, a nasal air stream is modulated, first by strengthening – weakening, then by weak glottal vibration, then by a strong voice burst, followed by a weaker one, of [m] colouring, and finally by mouth opening and a voice burst with schwa resonance. This results in a double iambic pattern, each with ascending pitch, a different rhythmicity from the one discussed in 3.1.

Laughing on a nasal air stream conveys somewhat subdued hilarity, a chuckle. In the present case, it occurs in preparation of an unrestrained oral laugh together with the dialogue partner.

Speaker mpi of **l06** also shows this difference between unrestrained and restrained laughter in figures 5 and 6 and in *audio_file_5-6.wav*.

The unrestrained laughter occurs right at the beginning of the dialogue after mpi has emphatically stressed the utter stupidity of the TV series, by saying *"ich hatte schon 'n bisschen vergessen, wie extrem unglaublich schwachsinnig die LINDENSTRASSE ist."* ("It had already somewhat slipped my mind how extremely unbelievably idiotic LINDENSTRASSE is."). He gives *"schwachsinnig"* a force accent [3]. (He is the speaker who has the highest number of force accents for negative emphasis in the whole corpus.) He then highlights his own characterization of the series by unrestrained laughter with a wide-open mouth.

Later-on, he reports on scenes that were presented to him in his video clip and refers to one by the non-word *"didelidu"*, which he again finds hilarious but is less emphatic about, so restrains his laughter to a chuckle by closing his mouth and modulating a nasal air stream.

Figure 4: f061_jm *"ne"* + nasal + oral laugh, voiceless, voiced: spectrogram, f0 (plain), energy (dotted)

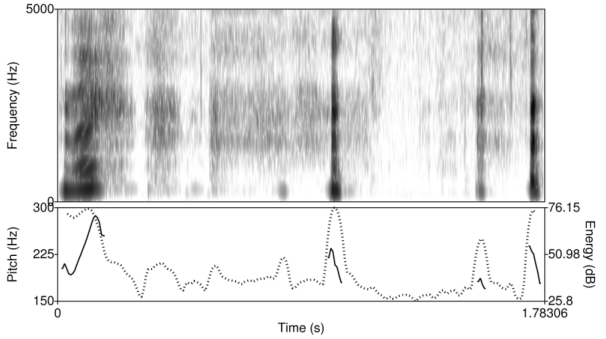


Figure 5: l06_mpi oral laugh, voiceless, voiced: spectrogram, f0 (plain), energy (dotted)

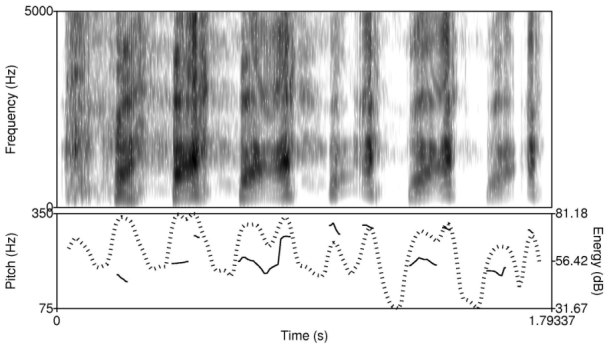
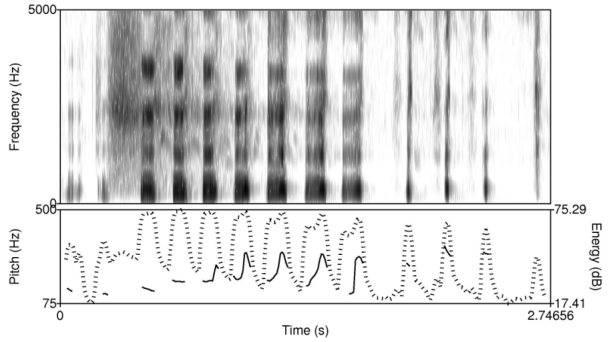


Figure 6: l06_mpi nasal laugh, voiceless, voiced: spectrogram, f0 (plain), energy (dotted)



3.4. Rhythmicity

The nasal or oral air stream is modulated by alternating phonation from glottal opening to various types of vibration and vice versa, and by imposing a dynamic and a duration structure especially on the voiced sections in a sequence. These modulations create rhythmic patterns. In fig. 3, we have seen a sequence of equidistant and equiprominent voice bursts, in fig. 4 an iambic pattern. In fig. 5, the rhythm is even more finely structured: an upbeat of 2 short voice bursts is followed by 4 longer double-peaked ones, grouped in twos of strong – weak energy, i.e. a trochaic pattern, which is clearly perceived in the signal.

In fig. 6, the first 4 voice bursts are evenly spaced and of equal energy on an ascending pitch scale. The next 3 form another block of still evenly spaced but longer bursts, of which the first 2 have well developed f0 upglides (perceivable as such), whereas in the third f0 jumps up abruptly from a low creaky section muffling the rising pitch movement. This together with the decreasing energy in this block creates a dactylic pattern. Then follows a third block of quite short and weaker voice bursts on a high rising pitch level, still evenly spaced.

3.5. Laughing interaction in dialogue

Laughing phenomena are not only sequenced and timed carefully within one speaker according to the communicative functions they are to fulfil but also as part of the interaction with another speaker in dialogue. In **f061**, mg makes an isolated laugh burst just before jm's utterance of fig. 2, seeing the funny side of the clash between task and execution, which jm is leading to. During jm's subdued laughter section in fig. 2, mg produces laughter, followed by speech-smile and then by speech-laugh on the utterance *"ich glaub, darum geht es hier nicht."* ("I don't think that's an issue here."), finally turning into laughter. Then jm agrees *"dann machen wir daraus ein viertägiges."* ("In that case we turn it into a 4-day meeting."), ending in a smile, followed by laughter, during which mg joins in with the speech-laugh and laughter of fig. 3. Towards the end of the latter, jm says *"gut. also machen wir das so."*, on which mg comments with a smiling *"ja, gut."* ("All right."). Then both speakers join in laughter finishing off the dialogue. The two speakers' laughing is coordinated action as part of their joint task-solving in the Appointment-making Scenario. This is illustrated by the complete *audio_files_7-jm* and *7-mg*, which are the two recorded channels of the dialogic interaction. They may be listened to, separately or conjointly, by opening them, for instance, in *cool edit.*

In the Video Task Scenario of **l06**, the situation is quite different. The speakers are not engaged in a joint task-solving goal, but simply talk about the differences they have observed in their respective video clips of the TV series. Speaker mpi sets the theme of emphatic evaluation which he embellishes with amusing wordings accompanied or followed by hilarious laughter. In this, he dominates the dialogue and stimulates speaker tra into laughing, which is, however, never so uproarious as his own.

The *audio_files_8-tra* and *8-mpi* provide examples of the two speakers tra and mpi mutually triggering laughter by facetious descriptions of the soap opera excerpts they have seen. tra gives an account of the scene where a gay chap makes advances to the Turkish doctor, and calls it *ein bisschen anbaggern* "a little digging", which sends mpi off into uproarious laughter, partially overlapping tra's continuation of his story, and his summarising comment: *du bist doch auch 'n kleines bisschen schwul, und so* "you are a little bit gay yourself, aren't you, and that jazz". *und so* is said tongue in cheek with a speech-smile running right through it. This gets mpi into a hilarious mood again because he did not have this episode in his video clip, and refers to the person who spliced the videos as withholding the juicy scene from him. He produces the long utterance *w<Z>as? solche Szenen hat <P> Benno mir gar nicht gezeigt* with a speech-smile throughout. It sends tra into hilarious laughter partially overlapping mpi's turn. So, here we have instances of a speech-smile developing into laughter, and vice versa, within one speaker, and laughing phenomena controlling the interaction between speakers.

In the following orthographic transcripts of these interchanges between jm and mg in **f061** and between tra and mpi in **l06**, sequential turns, with the speakers' IDs, are numbered from 01. Partial overlap is symbolized by giving the two speakers' turns the same sequential number. Overlays of speech-smile and speech-laugh are annotated by enclosing the stretch of overlaid speech in <: :>; <P>, <A> = pause, breathing, <Z> = hesitation.

f061 *audio_files_7-jm* and *7-mg*

jm01 und im Dezem=/- <A>
mg01 ja, wenn wir den zweiten und dritten Dezember <P>
mg02 genommen hatten,
jm02 ja, das würde gehen. <P>
mg03 und dann müssen wir eben dann<Z> daraus 'nen viertägigen<Z> Verabredung machen. <P>
jm04 das wären ja zweimal zwei Tage hintereinander. <A>
mg05 ja, wenn wir keinen anderen Termin finden. <P>
jm06 ja, das würde natürlich gehen. aber das ist bestimmt nicht so<Z>
mg07 <laughter>
jm07 <:<speech-laugh>glücklich:> <laughter>. <A>
mg08 <:<speech-smile> ich glaub', darum geht es hier:> <:<speech-laugh> nicht:>
mg09 <laughter>
jm09 gut, also <P> machen wir das so. <A>
mg10 ja. <A>
jm11 denn machen wir daraus ein <:<speech-smile>Viertägiges:>.
jm12 <laughter>
mg12 <:<speech-laugh> ja:> <laughter>.
jm13 <A> ja, dann hätten wir 's, nicht?
mg14 <:<speech-smile> ja, gut:> <laughter>.
jm15 <laughter>
mg15 <laughter>

l06 *audio_files_8-tra* and *8-mpi*

tra01 das ist +/der<Z>/+ der Türke, der auch +/in der/+ in der<Z>/+ in dem Haus wohnt, <A> dieser türkische Doktor.
mpi02 <äh>
tra03 <A> dass er den irgendwie anbaggert?
mpi04 <äh> null Komma null gesehen.
tra05 <A> mhm, das war nämlich irgendwie die zweite Partyszene, die bei mir irgendwann auftauchte so, wo +/v=/+ <äh> <P> der ihn so 'n bisschen anbaggert
mpi06 <laughter> <:<speech-laugh> ah, ej:>
tra06 und meint +/du bist/+ <A> du bist doch auch ‘n kleines bisschen schwul <:<speech-smile> und so:> +/und/+ und <häs>
mpi07 <:<speech-smile> w<Z>as? solche Szenen hat <P> Benno mir gar nicht gezeigt:>, Alter, hat er mir vorenthalten.
tra07 <laughter>
tra08 ja, woraufhin der Türke natürlich die Party irgendwie beleidigt verlässt.

4. OUTLOOK

Starting from sampled instances of three types of laughing phenomena – laughter, speech-laugh, and speech-smile – this paper has looked at their phonetic patterning and communicative function in interactive speech processes, considering laughter pulmonic air stream modulation in its own right, in alternation with, or superimposed on, the air stream modulation in the speech channel. Even such a small database suggests that fine phonetic detail of laughing is highly structured in its link with dialogic interaction. Its fine-grained analysis in instances of communicative settings can provide insights into the acoustic make-up, including rhythmic factors, as well as into the pragmatics of laughing in interaction. The auditory and acoustico-graphic approach advocated here needs to be extended to a much broader database of spontaneous speech from various scenarios. The investigation will also have to consider to what extent the phenomena are determined by the language and by the individual speaker.

5. REFERENCES

[1] Abercrombie, D. 1967. *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

[2] Bachorowski, J.-A., Smoski, M. J., Owren, M. J. 2001. The acoustic feature of human laughter. *J. Acoust. Soc. Am.* 110, 1581-1591.

[3] Kohler, K. J. 2005. Form and function of non-pitch accents. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK)* 35a, 97-123.

[4] Kohler, K. J., Peters, B., Scheffers, M. 2006. The Kiel Corpus of Spontaneous Speech. Vol. IV. German: VIDEO TASK SCENARIO (Kiel-DVD #1). pdf file in *Sound Patterns of German Spontaneous Speech*: www.ipds.uni-kiel.de/kjk/forschung/lautmuster.en.html

[5] Ladefoged, P. 1967. *Three Areas of Experimental Phonetics*. London: Oxford University Press.

[6] Nwokah, E.E., Hsu, H.-C., Davies, P. & Fogel, A. 1999. The integration of laughter and speech in vocal communication: a dynamic systems perspective. *J of Speech, Lang & Hearing Res*, 42, 880-894.

[7] Pätzold, M. and Simpson, A. P. 1995. An acoustic analysis of hesitation particles in German. In *Proc. XIIIth ICPhS*, Stockholm, 512-515.

[8] Peters, B. 2001. 'Video Task' oder 'Daily Soap Szenario' - Ein neues Verfahren zur kontrollierten Elizitation von Spontansprache. *Manuscript.* http://www.ipds.uni-kiel.de/pub_exx/bp2001_1/Linda21.html

[9] Pike, K. L. 1943. *Phonetics.* Ann Arbor: The University of Michigan Press.

[10] Sundaram, S., Narayanan, S. 2007. Automatic acoustic synthesis of human-like laughter. *J. Acoust. Soc. Am.* 121, 527-535.

[11] Trouvain, J. 2003. Segmenting phonetic units in laughter. *Proc. XVth ICPhS* Barcelona, 2793-2796.

[12] Trouvain, J. 2001. Phonetic aspects of “speech-laughs”. *Proc. 2nd Conference on Orality & Gestuality (ORAGE)* Aix-en-Provence, 634-639.

PERCEPTION OF SMILED FRENCH SPEECH BY NATIVE VS. NON-NATIVE LISTENERS: A PILOT STUDY

Caroline Émond[1], Jürgen Trouvain[2] and Lucie Ménard[1]

[1]Université du Québec à Montréal and [2]Universität des Saarlandes
caroemond@hotmail.com, trouvain@coli.uni-sb.de, menard.lucie@uqam.ca

ABSTRACT

Smiling is a visible expression and it has been shown that it is audible too (Tartter [5], Tartter and Braun [6], Schröder et al. [4], Aubergé and Cathiard [1]). The aim of this paper is to investigate the perception of the prosody of smile in 2 different languages. In order to elicitate smiled speech, 6 speakers of Québec French were required to read sentences displayed with or without caricatures. The sentences produced were used as stimuli for a perception test administered to 10 listeners of Québec French and 10 listeners of German who had no knowledge of French. Results suggest that some prosodic cues are universals and others are culture specific.

Keywords: prosody, emotion, smiling, smiled speech.

1. INTRODUCTION

It has been hypothesized by Tartter [5] that because the vocal tract is altered from its neutral position when there is smiling, there should be an audible effect. Even when neutral and smiled speech is produced in controlled and unnatural conditions (mechanically isolated speech samples – one syllable nonsense words), results showed that smiling has audible effects and those effects are automatically associated with positive emotions. Moreover naive listeners can reliably identify smiled speech as such.

Like Tartter [5] and Tartter and Braun [6], Schröder et al. [4] showed in the first perceptual experiment of their study that listeners can accurately discriminate utterances produced with a mechanical smile when presented with their neutral counterparts in audio condition. According to Ekman et al. [2], amusement smiles (so-called Duchenne smiles, i.e. produced with upturned mouth corners, raised cheeks, and crinkling of eyes) and smiles without any relation to amusement are different. Following them, Schröder et al. [4] showed in a second perceptual experiment that spontaneous stimuli (vs. mechanical stimuli) are discriminated just as the amused ones in audio conditions.

In Aubergé and Cathiard [1], the visual and audiovisual conditions were analyzed in relationship with the audio one. They found that the audio modality contained a lot of information in such a visible emotion as amusement. This information is not due only to the change of the vocal tract from its neutral position but there is also "a specific manipulation of the prosody of the speech" (p. 96) in the expression of amusement.

According to those references, it seems universal that smile in speech can be perceived by listeners. The objective of the present study is thus to compare the perception of smiled French speech across 2 languages: Québec French (QC) and German (GE).

2. METHOD

Before going any further, it is important to note that this pilot study is part of a larger study focusing on the prosodic correlates of smiled speech in QC (Émond [3]).

2.1. Corpus, participants and recordings

For the production part of the study, 10 humourous caricatures published in daily newspapers (*La Presse*) were chosen in order to elicitate smiles. 30 fillers were added (20 sentences presented alone or with drawings and the titles of the caricatures without the drawings (10), n = 40).

6 participants ranging in age from 22 to 34 years old (3 men, 3 women) with QC as L1 were recruited in the department of linguistics at the Université du Québec à Montréal. They were not aware of the study's true objective before the recordings. Stimuli were semi-randomized across the speakers, the first ten utterances being neutral.

Speakers were audio-video recorded. The recordings took place in a sound proof room with the following material: an IBM laptop, a Panasonic AG-DVC30 numeric camera, a (DAT) Tascam

numeric recorder and a Shure Beta 58A dynamic microphone.

The participants sat in front of the laptop, with the microphone about 30 cm from their mouth. The instructions were first presented on the screen and a training phase preceded the task. They were asked to read 40 sentences out loud and the test lasted about 10 minutes.

2.2. Selection of the test corpus

The data were digitized with Adobe Premiere and segmented with Goldwave. First, the utterances deemed as spontaneous smiled speech were selected. Audiovisual inspection was done to ensure those utterances were produced with the Duchenne effect. 32 utterances (out of 240 – 40 sentences x 6 speakers) were selected and to these were added 12 fillers i.e. utterances perceived as neutral (n = 44). It is important to note that the preselection of the corpus was only made to select a subset of sentences to be submitted as a perceptual experiment. Even though produced smiled speech corresponds here to the sentences produced with the Duchenne effect, a sentence will be said to be smiled speech only if it is perceived as such by the listeners. Our method is thus clearly listener-oriented, in part because of the origin of our listeners.

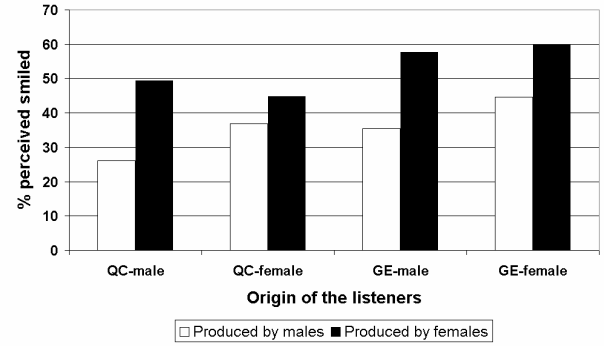
2.3. Perception test

The subset of sentences described above was used for an auditory perceptual experiment. 20 participants aged from 20 to 39 years old participated to the perceptual test: 10 Germans (GE – 5 men, 5 women) and 10 Quebecers (QC – 5 men, 5 women). Stimuli were presented via PC loudspeakers in a random order and mixed condition. The test took place in a quiet room and lasted about 5 minutes. The task consisted of a forced choice between 2 possible answers: smile or not smile (neutral).

3. RESULTS

On the whole, all the listeners behaved the same i.e. the number of utterances perceived as smiled by the GE is proportionally the same as the QC even if only the latter benefit from lexical and semantic access. Indeed, since no delexicalization method was used, QC listeners could have used segmental and prosodic cues, as well as some semantic content. On the contrary, for the GE listeners, the QC sentences did not have any semantic content and can be compared to some kind of "ecological delexicalization." It can thus be hypothesized that GE listeners refered only to prosodic and phonetic parameters. We shall come back to this issue later. Fig. 1 shows the distribution of the utterances perceived as smiled by both linguistic groups. This figure shows that the utterances produced by QC females are perceived more smiled than the ones produced by QC males by all the listeners. However, GE listeners tend to perceive a larger percentage of the sentences as smiled, compared to QC listeners.

Fig. 1: Perception of smiled speech based on the origin of the listeners.



In order to further investigate between-group differences, fig. 2 to 5 represent the perception of smiled speech by linguistic group and gender.

The perception of QC speakers is presented in fig. 2 (male listeners) and fig. 3 (female listeners). Data are grouped according to the origin of the listeners (light bars = QC listeners; dark bars = GE listeners). If we compare for example the perception of the utterance F_Am_28 by male and female listeners (fig. 2 & 3) it can be seen that 3 QC females (out of 5), 3 QC males (out of 5), 1 GE female (out of 5), and 4 GE males (out of 5) perceived this sentence as smiled. Those results show that there is a difference of perception between both linguistic groups but also between gender for the GE listeners.

Concerning utterance F_Lu_35 (fig. 2 & 3), 5 QC females, 3 QC males and all the GE listeners perceived it as smiled. This result suggests the presence of universal prosodic cues but that the perception differs in the members of the same cultural community.

Fig. 2: Recognition rates; female speakers; male listeners.

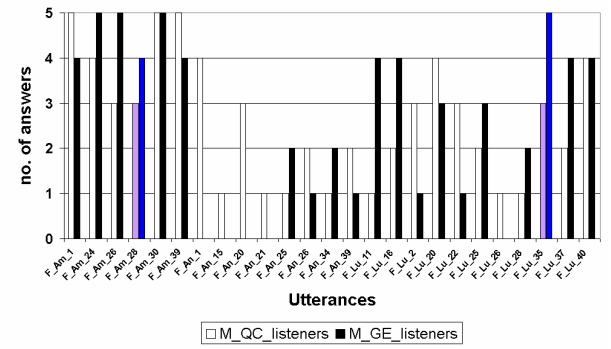


Fig. 3: Recognition rates; female speakers; female listeners.

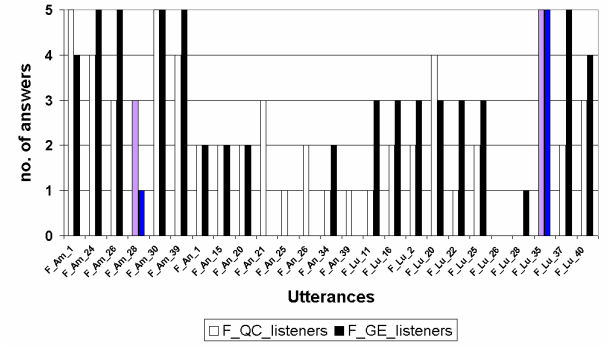


Fig. 4: Recognition rates; male speakers; male listeners.

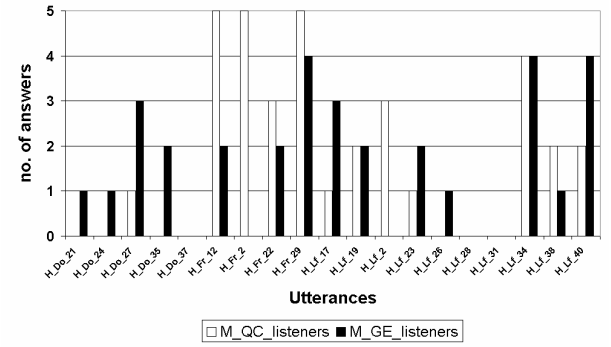
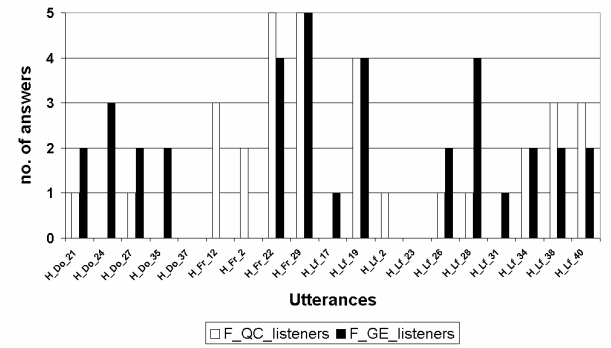
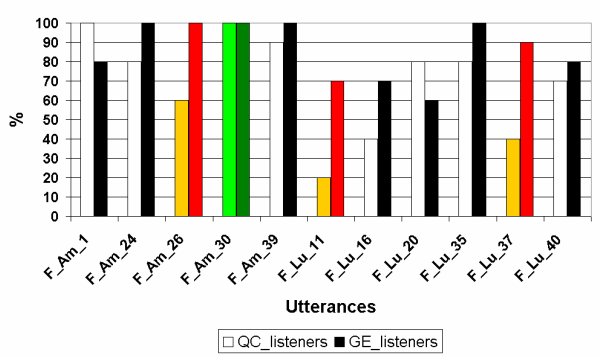


Fig. 5: Recognition rates; male speakers; female listeners.



Because of the important variability in the identification of smiled speech across stimuli, a subset of the corpus was used to study possible prosodic correlates of perceived smile. Utterances perceived as smiled by 70% of the listeners were thus considered. Fig. 6 shows 11 utterances out of 25 for 2 female speakers. There is one clear case where 100% of the listeners agree (F_Am_30). There is disagreement for a couple of utterances (F_Am_26, F_Lu_11, F_Lu_37) where more GE listeners found they were smiled compared to the QC listeners.

Fig. 6: Recognition rates of at least 70% in one of the listener group of 2 QC female speakers.



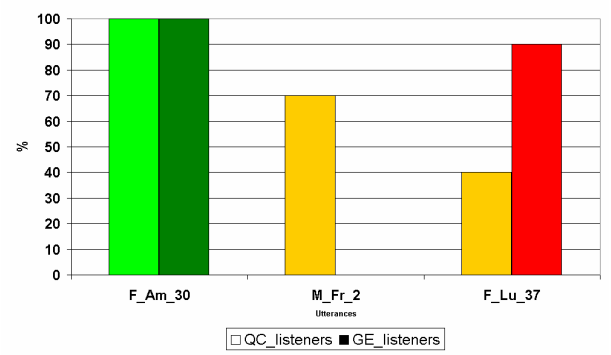
Concerning utterances produced by males, one sentence was perceived as smiled by at least 70% of the listeners.

This suggests that some universal cues may be present in the sentences for which most of the listeners agree, whereas culture-specific prosodic cues are produced in the sentences for which there is disagreement between the listeners. It is interesting to note that only 3 speakers out of 6 seem to be "more smiley." This may indicate that even if the utterances were presented in a random order in a mixed condition, there could be a speaker effect.

An unexpected phenomenon happened during the test: the spreading of the smile from an utterance to another. For that reason the neutral counterpart of the perceived smiled sentences were not present in the test. Maybe adding more fillers between the utterances would have helped. We examined the pitch range (max – min) and shape of the F_0 curves (with the standard software Praat) just to see if there could be an indication in the disagreement of the perception by both linguistic groups, if there could be a path to follow in further investigations even if we do not have those

utterances. Then we had a look at few cases for which all the listeners agreed (F_Am_30) and for which there was strong disagreement between the listeners (M_Fr_2, F_Lu_37). For the sake of clarity, the perception scores are depicted in Fig. 7 for the three sentences. We are aware that no comparison can be made at this stage but we observed a narrower pitch range for M_Fr_2 compared to other speakers which brings the idea that pitch range can be responsible for more subtle and perhaps sometimes culture-specific prosodic cues.

Fig. 7: Three types of recognition of smiled speech: full agreement across both groups of listeners (left), strong disagreement of the GE listeners (mid) and of the QC listeners (right).



4. CONCLUSION

We have seen that basically all the listeners behave the same i.e. the number of utterances perceived as smiled is nearly the same for both linguistic groups. The cases where all the listeners agreed may lend support the hypothesis of the existence of universal prosodic cues for the perception of emotions. However, strong disagreements sometimes arose. We can believe that this hypothesis is also supported when most of GE listeners perceive an utterance as smiled where most of QC listeners did not (e.g. F_Lu_37, fig. 6). In which case, we can suppose that pragmatic and lexical content play a strategic role. Following Thompson and Balkwill [7] our results suggest that "emotions are communicated prosodically through a combination of culturally determined and universal cues" (p. 421). In other words the recognition of smiling in speech is not as universal as expected.

We showed also that there are cases where the difference of the perception is due to listeners' gender. What is suggested here is the perception of emotion differs depending on speakers' and listener's gender. This criterion needs to be taken into account for any research in the field of emotions. For future work, it would necessary to have all the neutral counterparts to the utterances perceived as smiled as well as more participants for the perception test. Listeners from another dialect of the language of the speakers (e.g. French from France vs. French from Québec) would be another interesting variable to study.

Finally, we should not forget that elicitating emotions in this area of research is always harder than expected because obviously we deal with human beings and imponderables are numerous and frequent. The relationship between the experimenter and the participants is crucial. So, the collecting of spontaneous data in an experimental context added to the idiosyncratic aspects of smile and laughter remain at this time a sizeable challenge to researchers, one that should be tackled in future projects.

5. REFERENCES

[1] Aubergé, V., Cathiard M. 2003. Can we hear the prosody of smile? *Speech Communication* 40, 87-97.

[2] Ekman, P., Davidson, R. J., Friesen, W. V. 1990. Duchenne's smile: emotional expression and brain physiology. *Journal of Personality and Social Psychology* 58, 342-353.

[3] Émond, C. 2006. Une analyse prosodique de la parole souriante : une étude préliminaire. *Actes des XXVI^e Journées d'étude sur la parole (JEP)*, Dinard, 147-150.

[4] Schröder, M., Aubergé, V., Cathiard, M. 1998. Can we hear smiles? *Proc. ICSLP Sydney*, 559-562.

[5] Tartter, V. C. 1980. Happy Talk: Perceptual and acoustic effects of smiling on speech. *Perception & Psycho-physics* 27 (1), 24-27.

[6] Tartter, V. C., Braun, D. 1994. Hearing smiles and frowns in normal and whisper registers. *Journal of Acoustical Society of America* 96 (4), 2101-2107.

[7] Thompson, W. F., Balkwill, L.-L. 2006. Decoding speech prosody in five languages. *Semiotica* 158–1/4, 407-424.

PSYCHOLOGICAL AND CROSS-CULTURAL EFFECTS ON LAUGHTER SOUND PRODUCTION

Marianna De Benedictis
marianna_de_benedictis@hotmail.com
Università di Bari

1. ABSTRACT

The research within this paper is intended to offer a longitudinal examination of the laugh signal, in order to gain a deeper understanding of how this complex phenomena happens and under which circumstances it can change. So that the main purpose was to identify all the possible sound patterns and elements of a laugh signal, i.e. also those features which are commonly ignored because they are not included among the most well-known *ahahah* sound pattern. Moreover, taking a case study approach, the research tried to question the possible differences existing between "spontaneous" and "intentional", as well as an Italian and a German laugh sound.

2. INTRODUCTION

The non-verbal vocal behaviour to express amusement and mirth can be as different as a laughter and a speech-laugh [1]. Nevertheless it has often been argued that the various ways of expression can change on the basis of the intention of the subjects [2], of their personality and culture [3] [4]. In this study these aspects will be taken into account, giving attention to the complex phonetic structure of the laugh sound as Bachorowski did [5].

2.1. AIMS OF THE STUDY

Starting from the assumption that the sample data was very limited, because the research was meant to be a pilot study, the following purposes were outlined considering the variety of the corpus collected:

a. to identify the possible cultural differences in laughter sounds;
b. to highlight the differences between a spontaneous and an aware laughter, i.e. how a person uses laughter purposely;
c. to analyse all those features of laughing (inspiration, vocalised nasalization) usually not taken into consideration in existing literature.

2.2. TERMINOLOGY

The study of laughter doesn't have a precise terminology, as already Trouvain [13] affirms, but the most common approach is to segment a laugh signal considering it as articulated speech. This can cause some problems because it implies the false assumption that all those aspects not included among the consonant-vowel pattern cannot be classified. So that the terminology, suggested by Bachorowski, revealed to be more suitable to include and to classify the following phonetic aspects:

- laugh call: a rhythmic laugh unit with the vocalic segment as the syllable nucleus preceded by a consonantal segment;
- glottal pulses;
- inhalation: it can be vocalized high pitched and/or characterised by a strong vocalized nasalization.

Aside of these there are also some particular features of laughter which will be examined:

- retained laughter: laughing sounds produced by voluntarily modifying the overflowing sound of the laughter, in the attempt to hide or restrain it. They are different from the "low-pitched chuckle" [5], because of the strong air irruption and often high-pitched vocalization;
- monosyllabic laughter: firstly identified by Edmonson [9]. They are composed by a single laugh call and are called also "comment laugh" [12].

3. DATA RECORDING

A corpus of laugh sounds was collected from three subjects (all female) of two different cultures German and Italian, in two different conditions: spontaneous and aware. The stimulus to elicit the laughter was different according to the spoken language. The spontaneous situation was possible because the two girls (D. and A. – see table 1) had been given the task to listen to the sketch in order to offer their own opinion on the sense of humour

found in the stimulus, the microphone was on while they were listening to the comic sketch with the earphones. In fact they didn't know of being recorded and that the purpose of being invited in the lab was to collect so many laugh signals as possible. The aware situation was due to the fact that the girl (M.) knew the aim of the research, i.e. the recording of laugh sounds. So that it seems reasonable to talk of "non-spontaneous laughter", not because they were forced, but they were not free of the so-called "self-presentation" bonds [6]. In both conditions the microphone was set at 20 cm from the mouth of the subjects.

Unlike other studies, the subjects were not alone in the laboratory, but they were with the author of the research (S.). As the laughter is essentially a social behaviour only using a movie as an solicitator cannot be satisfactory [7] and [8].The experimenter (female) accompanied all the three subjects and she was supposed to interact with them but avoiding to overlap their laughter (as matter of fact her voice remains in the background of the recordings). Nevertheless her laugh and speech production were slightly audible in the recording and will allow some consideration and comparison.

Table 1 – Laughter sample pool

	D.	A.	M.	S.
time of recording	3 min. 15 sec.	5 min. 45 sec.	5 min. 45 sec.	9min.
n ° laughs	15	9	14	4

The recordings were than digitalized with the software WASP and segmented into smaller pieces lasting less than 3 seconds.
Within this sample pool all the isolated laugh sounds and speech-laughs were saved separately.

Table 2 – Subject recorded

	D	A	M	S
age	23	26	36	26
culture	German	Italian	Italian	Italian
situation	unaware	unaware	aware	aware

4. THE RESULTS

Although generally there is a great concordance with other studies, it was possible to analyse exceptions, or "uncommon" phenomena, the investigation of which can contribute to the understanding of the laughter in its complex and articulated nature. Consequently it is necessary to detach from those studies which try to limit it in its most stereotyped and common aspects (*ah ah ah* sound). It was recognized that Provine's limitation to stereotypical involuntary laughter covers the domain inadequately and that across natural languages the orthography for representing stereotypical laughter in the written mode is not the same as has been already mentioned by Trouvain [13].

The most relevant aspects found were as follows:

- retained laughter
- monosyllabic laughter

The *retained* laughter is a very important example of voluntary modification of laughing, in the attempt to hide the audible laugh expression. Although Backorowski noted that there are also low-pitched laugh signals, it is important not to confuse those ones with the retained laughter, whose features were already described by Darwin [10]:

- a strong inhalation obstructs the normal expiratory process of laughter (circled in fig. 1 and 2).
- the closure of the mouth prevents the sound exiting, so that the air exits through the nose producing a strong audible nasal expiration.
- control of the movement of the glottis, emitting very short pulses of low intensity.

In the following examples (figure 1 and 2) one can notice the complete irregularity of the laugh signal, in contrast with the well-known rhythmic sequence of laugh calls with the typical aspiration phase. Here it is clear the effort of both subjects to control the otherwise bondless emission of vocalization.

Figure 1 – Example of retained laughter in S.

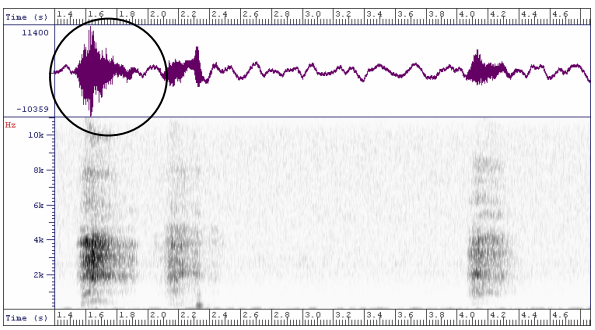
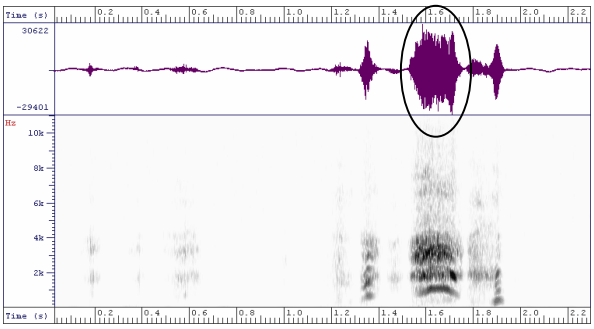


Figure 2 – Example of retained laughter in A.



Unfortunately this kind of laughter could be only identified with the perception of the experimenter, because there were no video recordings, so that no clear visual cues were available (covering the mouth with hand, avoiding eye contact with social partner).

The *monosyllabic* laughter was considered as composed of a single laugh call sandwiched with silence (2 sec.) and without overlapping acoustic events. The examination of this kind of laugh emission revealed that the laugh sound produced by a person is very similar both in long laugh episodes and monosyllabic ones. As a matter of fact, comparing the spectrum of a laugh call of a long laughter with that of a monosyllabic interesting similarities were visible (see figure 3 and 4, a monosyllabic and a laugh episode of subject A.). So one can hypothesize that, apart from the number of laugh syllables emitted, the movement of the glottis within the subject remains more or less the same. Subject A. and D. emitted this kind of monosyllabic laughter, or comment laugh, while the subject M. didn't. Probably the difference can be explained with the fact that this latter was aware, so that she supposed that the kind of laughter needed were those song-like and acoustically well identifiable, instead of small expiration or movement of the glottis.

Fig. 5 presents an example of monosyllabic laughter in subject D., while figure 6 gives an example of a longer laugh produced by the same subject. Comparing the laugh syllable in fig. 5 and fig. 6, one can clearly see that the sound produced during a monosyllabic laughter is similar to the one of the laugh bout. In both cases it is evident that the glottal movement is mainly of complete closure, instead of aspiration (as it was the case of the subject A.)

Figure 3 – Example of mono-syllabic laughter in A.

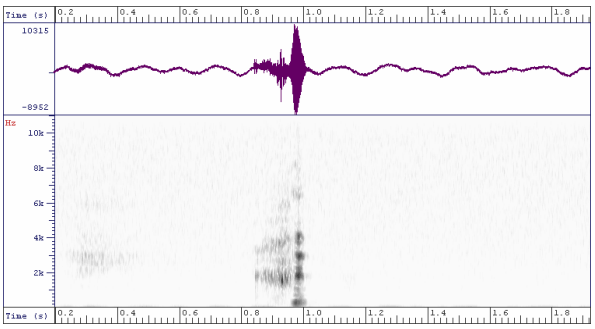
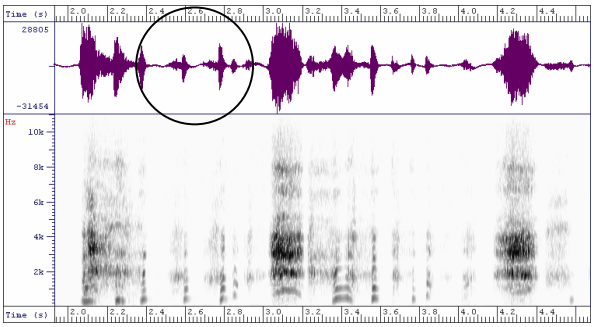


Figure 4 – Example of laugh episode in A.



Moreover, one can argue that the monosyllabic laughter in subject D. (fig. 5) is completely different from the one produced by subject A. (fig. 3), in which an aspiration phase precedes the vocalic emission. On the contrary here it is a strong glottal pulse.

Figure 5 – Example of monosyllabic laughter in D.

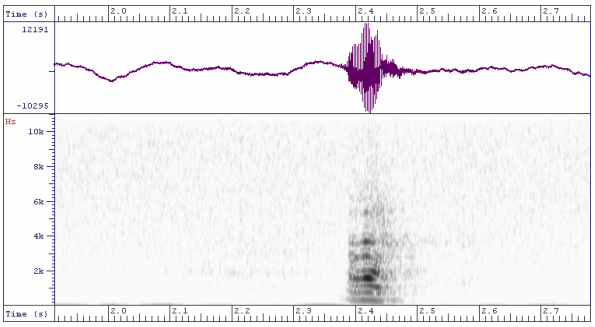
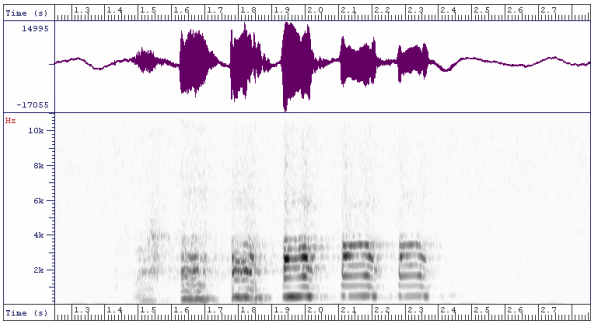
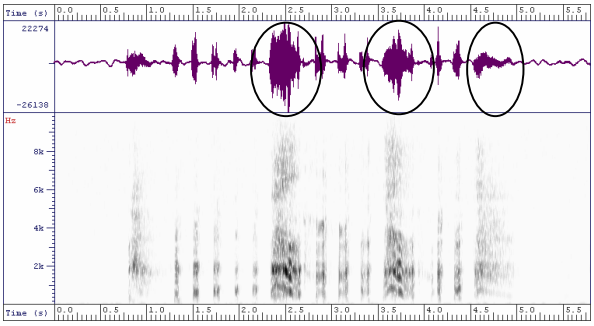


Figure 6 – Example of laugh bout in D.



Aside of this aspect, it can be considered how the laugh call is not composed necessarily by an aspiration phase followed by a vocalic segment, but just of a vocalic segment started with a sudden opening of the glottis.
Another aspect is the different use of the *vocalised inhalation* between the spontaneous and intentional laughter. The German subject made a rare use of this, compared with the two Italian subjects. But the subject M. made a greater use of them compared to A., most probably because it is easier to inhale letting the vocal cords vibrate than to emit the typical vocalised high-pitched sound of laughter, i.e. exhale as strong as the laughter emission is. Indeed it has been pointed out that the laughter expiration happens at a very low lung volume in which involuntary muscles movements are activated. So that it can be roughly assumed that the intention of producing loud laughing sounds was achieved thanks to the vocalized inhalation, which allowed the emission of vocalised *ahahah* pattern sounds. For example in the following figure it is possible to notice even three vocalized inhalation (circled in fig. 7).

Figure 7 – Example of laugh episode in M. with many vocalized inspirations



Furthermore, aside from all the consideration another interesting element was found, i.e. what Darwin [10] already pointed out. A laughing sound can be frequently confused with a crying one, especially was the case of those laughter rich in inhalation and pauses (file *laughter crying.wav*) can be confused with the sound produced when sobbing. Nevertheless, also in this case there isn't a relevant evaluation and evidence but the perception of the experimenter. Of course it would be required the realization of a perceptual test.

5. CONCLUSIONS

Given these results it can be noticed how the study of laughter needs a very complex way of analysis because there are many aspects to be considered. Consequently the one-dimensional approach which takes into account only the stereotypical sound of laughter [11] should be avoided, and the segmental portion should be extended to a variety of other phenomena like vocalized inhalation, nasal sounds and glottal pulses.

It was revealed that the use of a vocalized inhalation can help the subject to communicate being amused more easily, because the production of very strong audible laugh sound requires a particular lung effort, which is possible only in real spontaneous laughter. But a listener's evaluation of the degree of enjoyment or amusement in the laugh sounds with and without vocalized inhalation would be necessary.
The retention of laughter itself can be done by closing the mouth and controlling the vocal cords so that the emission of sound is reduced to short strong glottal pulses. It was found that a laughter does not necessarily start with an exhalation phase but also with a strong vocalized inspiration because of the contrasting force applied to invert the normal ongoing of the expression.
Relevant cross-cultural differences were not found except the rare use of vocalized inhalation within the German subject (D.). On the contrary they were very common in both the aware and unaware Italian subjects (A. M.). However, the sample size was much too small to draw any conclusions about such differences. So that any differences are found among these participants, it is impossible to know if these are merely the sorts of inter-individual differences found between any individuals within culture, or if they are due to the differences nationality and "awareness".

REFERENCES

[1] Trouvain, J. 2001. Phonetic aspects of speech-laughs, in *Proceedings Conference on Orality and Gestuality (Orage)*, Aix-en-Provence, 634-639

[2] Grammer, K. 1990. Strangers meet: laughter and non-verbal signs of interest in opposite encounters, *Journal of Nonverbal Behaviour 14*, 209-236

[3] Eibl-Eibesfeldt, Ir. 1974. Somiglianze e differenze interculturali tra movimenti espressivi In: Robert A. Hinde (ed), *La comunicazione non-verbale*. introd. of Tullio De Mauro, translation of R. Simone, Roma, Bari: 395-418

[4] Rothgaenger, H.; Hauser, G.; Cappellini, A.C. Guidotti, A. 1998. Analysis of laughter and speech sounds in Italian and German students, *Naturwissenschaft 85*, 394-402
[5] Bachorowski, J.A., Smoski, M. J., Owren M.J. 2001. The acoustic features of human laughter, *Journal of the Acoustical Society of America 110 (3)*, 1581-1597
[6] Scherer, K. 1986. Vocal affect expression: a review and a model for future research, *Psychological bulletin, vol. 99 (2)*, 143-165
[7] Martin, R. A. and Kuiper, N.A. 1999. Daily occurrence of laughter: relationships with age, gender and type A personality, *Humor, 12 (4)*, 355-384
[8] Provine, R. 2000. *Laughter: A Scientific Investigation.* London: Faber & Faber
[9] Edmonson, M., 1987. Notes on laughter, *Anthropological linguistics, 29 (1)*, 23-34
[10] Darwin, Ch. 1872. *The Expression of Emotions in Man and Animals*. London: Murray, (3rd ed. Paul Ekman, New York: Oxford Univ. Press, 1998) chapt. 8, 195-218
[11] Provine, R.; Yong, Y.L. 1991. Laughter a stereotyped human vocalization, *Ethology, 89 (115)*, 115-124
[12] Nwokah, E.E., Davies, P., Islam, A. Hsu, HC. & Fogel, A. 1993. Vocal effect in 3-year-olds – A quantitative acoustic analysis of child laughter. *Journal of the Acoustical Society of America*, 94 (6), pp. 3076-3090
[13] Trouvain, J. 2003. Segmenting Phonetic Units in Laughter. 15th ICPhS Barcelona.

Positive and Negative emotional states behind the laughs in spontaneous spoken dialogs

Laurence Devillers, Laurence Vidrascu

LIMSI-CNRS
91403 Orsay cedex
{devil, vidrascu}@limsi.fr

ABSTRACT

This paper deals with a study of laughs in spontaneous speech. We explore the positive and negative valence of laughter in the global aim of the detection of emotional behaviour in speech. It is particularly useful to illustrate the auditory perception of the acoustic features of laughter where its facial expression (smile type) is not visible. A perceptive test has shown that subjects are able to make the distinction between a positive and a negative laugh in our spontaneous corpus. A first conclusion of the acoustic analysis is that unvoiced laughs are more perceived as negative and voiced segments as positive, which is not surprising.

1. INTRODUCTION

Laughter is a universal and prominent feature of human communication. There is no reported culture where laughter is not found. Laughter is expressed by a combination of speech and facial expressions. In our study, only the audio channel is used. The laugh plays an important role in human social interactions and relationships. Laughs colour speech, they can be spontaneous (uncontrolled) or controlled with a communicative goal. Laughs represent a broad class of sounds with relatively distinct subtypes, each of which may function somewhat differently in a social interaction [1].

This paper deals with the study of laughs in a corpus of human-human dialogs recorded in a French medical call center [4]. Our global aim is the detection of emotion. In [4], laughter feature were used as a linguistic mark (presence or absence of this feature) in emotion detection system.

The majority of laughs in our corpus overlap speech, instead of cutting it. Few works investigate speech with simultaneous laughter; Nwokah [6] gives evidence that up to 50% of laughs in conversations between mother and child (English) overlap speech, Trouvain [12] reports that 60% of all labelled laughs in the German "KielCorpus of Spontaneous Speech" are instances which overlap speech. The so-called "speech-laughs" are around 58% of all the laughs in our French corpus of spontaneous dialogs between a caller and an agent. Our findings agree with Trouvain's and Nwokah's studies and contrast with Provine [7] who reported that laughter almost never co-occurs with speech. Acoustic manifestations of "speech-laughs" are much more variable and complex that isolated laughs.

Negative forms of laughter are also used in everyday communication. Some studies report that laughter can express negative feelings and attitudes such as contempt [9] and it can also be found in sadness [11]. There is evidence that gelotophobics have difficulties to distinguish between positive and negative forms of laughter. The concept of Gelotophobia can be defined [8] as the pathological fear to appear to social partners as a ridiculous object or simply as the fear of being laughed at. A typical symptom of Gelotophobia is the systematically attribution of (even innocent) laughter as being negative. In the context of spontaneous children speech, other examples of complex positive and negative laughs are laugh-cry, that is, crying that switches to laughter and back and forth and half-cry/half-laugh which is a combination of simultaneous laugh and cry.

In our corpus, there are a lot of negative contexts where laughs have different semantic meanings. The current study aims to analyse the characteristics of laughter in order to define negative laughs and positive laughs. We explore laughter manifestations occurring in our corpus using prosodic cues such as pitch and energy variation and duration measures. We assume, like [9], that the non-verbal events, also called "affect bursts", such as laughs or tears, are among the relevant features for characterizing an emotion. A closer analysis of the acoustic laughter expressed in this study has shown that this cue can be linked to positive emotional behaviour such as "I'm pleased, I'm relief" or negative emotional behaviour "I'm anxious, I'm embarrassed."

First, this paper reports on a perceptive test allowing the annotation of the valence of laughter and proposes a typology of laughter. Then it presents a first prosodic feature analysis of laughs.

2. CORPUS

The study reported in this paper is done on a corpus of naturally-occurring dialogs recorded in a real-life call

center. The transcribed corpus contains about 20 hours of data. The service center can be reached 24 hours a day, 7 days a week. Its aim is to offer medical advice. An agent follows a precise, predefined strategy during the interaction to efficiently acquire important information. His role is to determine the call topic and to obtain sufficient details about this situation so as to be able to evaluate the call emergency and to take a decision. The call topic is classified as emergency situation, medical help, demand for medical information, or finding a doctor. In the case of emergency calls, the patients often express stress, pain, fear of being sick or even real panic. The caller may be the patient or a third person (a family member, friend, colleague, caregiver, etc). This study is based on a 20-hour subset comprised of 688 agent-client dialogs (7 different agents, 784 clients). About 10% of speech data is not transcribed since there is heavily overlapping speech. The use of these data carefully respected ethical conventions and agreements ensuring the anonymity of the callers, the privacy of personal information and the non-diffusion of the corpus and annotations. The corpus is hand-transcribed and includes additional markings for microphone noise and human produce non-speech sounds (speech, laugh, tears, clearing throat, etc.). The corpus contains a lot of negative emotional behavior. There are more tears than laughs. The laughs are extracted from annotations in the human—generated corpus transcripts. For each segment containing a laughter annotation, the segment is labelled as laugher. With more than half of the cases in our corpus, the laughs are associated with negative feelings.
Table 1 summarizes the characteristics of the global corpus and the sub-corpus of laughs. Table 2 gives the repartition of non-speech sounds on the corpus.

Table 1. The global and laughter corpus:

	Global corpus	Laughter corpus
#callers	784(271M,513F)	82 (28M, 54F)
#agents	7 (3M, 4F)	4 (3M, 1F)
#dialogs	688	82
#segments	34000	119
Size	20 hours	30 mn

Table 2: Number of the main non-speech sounds markings on 20 hours spontaneous speech.

#laugh	119
#tear	182
# « heu » (filler pause)	7347
#mouth noise	4500
#breath	243

3. PERCEPTIVE TEST

3.1. Experimental corpus

In order to validate the presence of negative and positive laughs and to characterize the related emotion type in our corpus, a perceptive test was carried out. In a first manual annotation phase, two expert annotators created emotional segments where they felt it was appropriate for all the speaker turns of the dialogs (the units were mainly the speaker turns). The expertise here is related to the knowledge of the data and the time passed for emotion definition and annotation (here one year). We evaluated the quality of these annotations on the global corpus using inter-coder agreement (kappa coefficient of 0.57 for callers, 0.35 for agents) and intra-coder agreement (85% agreement) measures, and correlation between different cues that are logically dependent, such as the valence dimension and the classes of negative or positive labels. A subset of 52 segments (balanced on three classes: positive, negative and ambiguous) was selected from the 119 segments containing laugh. These segments were extracted from 49 dialogs between 49 callers (15 M, 34F) and 2 different agents (1M, 1F). 18 of these segments were previously labelled as positive, 18 as negative and 16 as ambiguous. The emotional segments were re-segmented in order to only keep the laugh with the smaller context allowing the same valence of the segments and keeping the privacy of the data. So, this experimental corpus was selected with mainly isolated laughs. Only 37% of the test segments are “laugh superposed to speech”. In the laughter corpus, the proportion of female voice is 68%. 12% of laughs are extracted from agent speaker turns, 88% from callers.

3.2. Protocol

20 French native subjects recruited in the LIMSI French laboratory (14M, 6F) had to listen the 52 segments and to decide of the valence: positive, negative or ambiguous and of the type of laugh. As said in [10], laughing is not only an expression of exhilaration and amusement, but is also used to mark irony or a malicious intention. An important function of laughter is social bonding. In our database, the laugh functions included social, amused and affective functions. 10 different types of laugh were proposed and defined (Table 3). These 10 types were obtained after data observations, annotation and discussions by three experts. The subjects also had the possibility to annotate two types of laugh per segment: one dominant and one in the background.

Table 3. Type of laughs

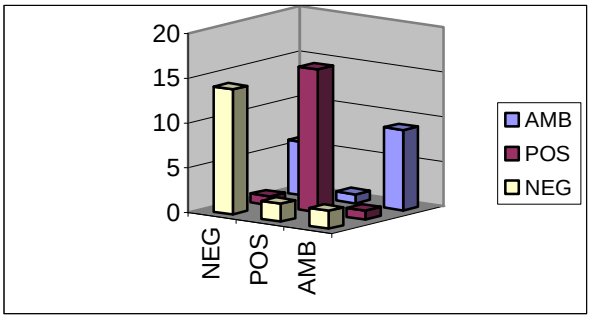
Positive labels	amused laugh joy laugh, sympathetic laugh

	polite laugh, relief laugh
Negative labels	disappointment laugh, embarrassed laugh, stressed laugh
Ambiguous labels	comment laugh ironical laugh

3.3. Results

In order to group the annotations of all the subjects per segment, two decision methods were used: a majority voting technique for the annotation of the valence (see Figure 1) and a soft vector representation [4] for the type of laugh (see Figure 2).

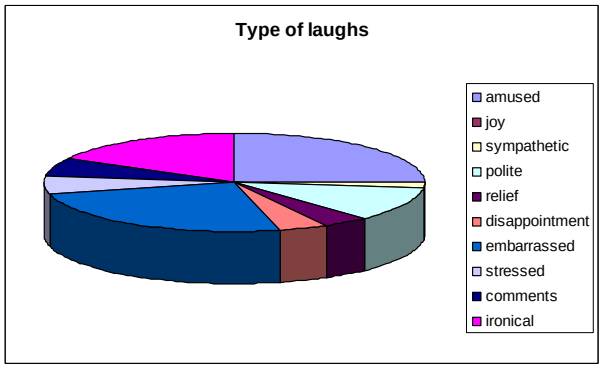
Figure 1. Repartition between the first annotations (abscissa) and the majority voting results on valence annotations of the perceptual test



The results in Figure 1 show that the positive laughs (88%) were easier to annotate than the negative laughs (77%). The ambiguous segments were judged mainly as ambiguous (56%) but also as negative laughs (37.5%).

The negative laughs perceived, linked to the mental state embarrassed is dominant for this corpus. For the positive laughs perceived, there are linked to the mental state amused.

Figure 2. Global repartition of the type of laughs obtained from the first coefficient of the soft-vector



4. PROSODIC FEATURES ANALYSIS

4.1. Features

A first prosodic analysis was carried out on positive/negative laughs and also on the four main types of laughs in our corpus: embarrassed (negative), amused (positive), ironical (ambiguous) and polite (positive). The laughter segments were exactly segmented on the laughs for this analysis. We computed some prosodic parameters using Praat software such as F0 statistics (mean, standard deviation), percent of unvoiced frames, energy (Pa^2s) and duration (ms). We used the "Lobanov" normalization for the F0 parameters.

4.2. Analysis

We can observe some trends on this small corpus. The main trends are that the energy and duration are higher for positive than for negative laughs and the percent of unvoiced frames is higher for negative than for positive laughs. When we looked more precisely at the four main types of laugh in the corpus, the trends were the following: the F0 measures are higher for amused laughs than for polite laughs; the duration is also highest for amused laugh, the percent of unvoiced frames is the highest and also the energy is the lowest for embarrassed and ironic laughs. These trends should be confirmed with a larger database.

5. CONCLUSIONS

This paper has addressed the less frequently discussed issue of negative laughter. The results of the perceptive test show that the subjects are perceptibly able to distinguish both laughs: positive and negative. A first prosodic analysis was carried out on positive/negative laughs and also on the four main types of laughs in our corpus. We have found some trends to characterize positive and negative laughs. These trends should be confirmed with a larger database. As a first conclusion, we can say that negative laughs are more unvoiced segments and positive laughs are more voiced segments what are not surprising [1].

6. REFERENCES

[1] Bachorowski, J-A, (1999) Vocal expression and perception of emotion, Current Direction in Psychological Sciences, 8, 53-57, 1999.

[2] Bachorowski, J-A, Owren, M.J. (2001). Not all laughs are alike: voice but not unvoiced laughter elicits positive affect in listeners. Psychological Science, 12, 252-257.

[3] Campbell, N, Kashioka, H., Ohara, R., (2005) "No Laughing Matter", Interspeech 2005.

[4] Devillers, L. Vidrascu, L., Lamel, L., (2005). "Challenges in real-life emotion annotation and

machine learning based detection", Neural Networks 18, pp. 407-422.
[5] Kennedy L., Ellis, D., 2004, "Laughter Detection in Meetings", NISTRT 2004.
[6] Nwokah, E.E, Hsu, H.C., Davies, P. & Fogel. A., 1999, "The integration of laughter and speech in vocal communication: a dynamic systems perspective." J of Speech, Lang & Hearing Res, 42, 880-894.
[7] Provine, R., 1993, "Laughter punctuates speech: linguistic, social and gender contexts of speech of laughter", Ethology, 95, 291-298.
[8] Ruch, W. & Proyer, R.T. (2005). Gelotophobia: A useful new concept?, 9th European Congress of Psychology, 3-8 July 2005, Granada.
[9] Schröder, M, 2000, "Experimental study of affect bursts", Proc. ISCA workshop "Speech and Emotion", Newcastle, Northern Ireland, 132-137.
[10] Schröder, M.: 2003, Experimental Study of Affect Bursts. Speech Communication, 40 (2003) 99-116
|11] Stibbard, R., 2000, "Automatic extraction of ToBi annotation data from the Readings/Leeds emotional speech corpus", Proc. ISCA workshop "Speech and Emotion", Newcastle, Northern Ireland.
[12]Trouvain, J., 2001, "Phonetic Aspects of "Speech-Laughs", conference on Orality and Gestuality ORAGE 2001, Aix-en-Provence, 634-639.

SOME PHONETIC NOTES ON EMOTION: LAUGHTER, INTERJECTIONS, AND WEEPING

Bernd Pompino-Marschall[1], *Sabine Kowal*[2] *& Daniel C. O'Connell*

[1]Humboldt-Universität zu Berlin, [2]Technische Universität Berlin
bernd.pompino-marschall@rz.hu-berlin.de, kowal-berlin@t-online.de,
doconnell@jesuits-mis.org

ABSTRACT

As consultant for the movie *My Fair Lady*, Peter Ladefoged must have found it a daunting task to resuscitate the phonetics of the early 20th century. The task in the present situation is just the opposite. We now have an extensive armory of equipment and methods; but the "primitive", marginal, and elusive element in the project is the subject matter itself: the phonetics of emotional expressions such as laughter, interjections, and weeping. All three of these areas have been marginalized and neglected in research in the scientific disciplines dealing with the communicative use of human language.

Keywords: emotional expressions, laughter, weeping, interjections

1. INTRODUCTION

Recent publications relevant to the phonetics of laughter include Partington [5], Trouvain [9, 10], and Vettin and Todt [11, 12]. An overview of related research on interjections is available in Kowal and O'Connell [2, 3]. And Hepburn [1] has engaged both in the transcription and analysis of weeping sequences.

It is our simple hypothesis that there are systematic phonetic differences among these emotional expressions, and between them and emotionally colored as well as "normal" non-emotional speech, particularly with respect to the onset level and course of f0.

2. MATERIAL AND METHOD

Our corpus consists of emotional outburst on the part of Mrs. Bennet (played by Alison Steadman) in one of the many motion-picture versions of Jane Austen's *Pride and Prejudice* (1995, BBC mini series). The locations of these utterances were isolated by a preliminary assessment and then analyzed in the Institute for German Language and Linguistics of the Humboldt University of Berlin.

Our preliminary findings are of Mrs. Bennet's emotions as expressed in the laughter sequences of the actress Alison Steadman. As part of the larger project on the phonetics of emotion (including laughter, weeping, interjections and other emotionally coloured speech), the bouts of laughter were analysed with the help of PRAAT regarding their durational call and segmental structure as well as their intonational realization.

3. RESULTS

The bouts showed an enormous range of variation with respect to all measured parameters, only partly dependent on the emotional meaning of the laughter. Accordingly, on the one hand, we find bouts signalling joyful surprise quite exaltedly with a three-to-five call structure (cf. Table 1) and throughout falling fundamental frequency in the highest register (up to 950 Hz and partly showing laryngeal whistle within their [h] segments).

Table 1: Example of the segmental make up of an exalted 4-call laugh bout (seg: segment; dur: duration in ms; f0: minimal and maximal f0 in Hz; cont: f0-contour – R: rise, F: Fall, s: short).

seg	aː	h	ɔ	h	ə	h	aʊ	h
dur	502	114	73	162	38	121	460	99
f0	833/ 959	906	843/ 883	870	795/ 865	830	683/ 804	
cont	RsF		RF		F		sRF	

At the other extreme, we find interjection like one-to-two call bouts at quite low f0, signalling nervousness (cf. Table 2).

Table 2: Example of the segmental make up of a nervous 3-call laugh bout (abbreviations as above).

seg	ə	h	a	h	ə
dur	78	84	100	89	77
f0			310		287
cont					

A wide range of laryngeal phenomena is observable: initial and final glottal stops, laryngalizations, diplophonia, octave jumps, and laryngeal whistle.

Furthermore, the laughter frequently combines with interjections or (partly ingressively voiced) breathing. Comparing the realisations of emotional sounds of the dubbed German version of the film, sometimes laughter can also be seen replaced by strongly emotionally coloured interjections.

4. DISCUSSION

In general, the laugh bouts analysed can be characterised as quasi 'nonarticulate' vocalisations with hightened laryngeal tension. Their call structure could thus be seen as a sequence of 'pressure syllables' ("Drucksilben") in the sense of Sievers [8] in contrast to the articulate syllable structure ("Schallsilben" according to Sievers).

Similar characteristics hold for weeping bouts.

In parallel to – and extending – Lindblom's H&H theory [4] we would like to propose classifying emotional vocalisations/speech – including laughter and weeping as well as non-'tame' interjections – along combined scales of 'tonus' as already proposed for interjections (cf. [6]): Normal speech thus ranges from hyperarticulate tonus of extremely careful articulation to relaxed hypo-speech. Hesitation vocalisations ("ehm") in non-'tame' production may even further reduce in articulatory tonus to more vegetative settings: Schwa-articulation followed by lip closing and velum lowering and devoicing.

Concerning laughter, we would state a hypo tonus for the supralaryngeal articulators and an extremely high tonus laryngeally. It seems to us that laughter and weeping as well as differentially 'tame' interjections (and other emotional vocalisation) might be classified in a systematic way along these lines.

5. REFERENCES

[1] Hepburn, A. 2004. Crying: Notes on Description, Transcription, and Interaction. *Research on Language and Social Interaction, 37,* 251-290.

[2] Kowal, S. & O'Connell, D.C. (eds.) 2003a. Interjektionen. *Zeitschrift für Semiotik, 26.*

[3] Kowal, S. & O'Connell, D. C. 2003b. Interjektionen im Gespräch. *Zeitschrift für Semiotik, 26,* 3-127.

[4] Lindblom, B. 1990. Explaining phonetic variation: a sketch of the H&H theory. In: Hardcastle, W.J. & A. Marchal (eds.) *Speech Production and Speech Modelling.* Dordrecht: Kluver. 403-239.

[5] Partington, A. 2006. *The linguistics of laughter: A corpus-assisted study of laughter talk.* London: Routledge.

[6] Pompino-Marschall, B. 2001. Connected speech processes as multitier/multiarticulator prosodic modulations. In: Puppel, S. & Demenko, G. (eds.), Prosody 2000, Posnan, 205-210.

[7] Pompino-Marschall, B. 2004. Zwischen Tierlaut und sprachlicher Artikulation: Zur Phonetik der Interjektionen. In: Kowal, S. & O'Connell, D.C. (eds.), Interjektionen. *Zeitschrift für Semiotik 26*, 71-84.

[8] Sievers, E. [5]1901. *Grundzüge der Phonetik.* Leipzig.

[9] Trouvain, J. 2001. Phonetic aspects of "speech-laughs". *Proceedings of the Conference on orality and gestuality ORAGE,* Aix-en-Provence, pp. 634-639.

[10] Trouvain, J. 2003. Segmenting phonetic units in laughter. *Proc. 15th ICPhS,* Barcelona, pp. 2793-2796.

[11] Vettin, J., & Todt, D. 2004. Laughter in conversation: Features of occurrence and acoustic structure. *Journal of Nonverbal Behavior, 28,* 93-115.

[12] Vettin, J., & Todt, D. 2005. Human laughter, social play, and play vocalizations of non-human prmates: An evolutionary approach. *Behaviour, 142,* 217-240.

IMITATING CONVERSATIONAL LAUGHTER WITH AN ARTICULATORY SPEECH SYNTHESIZER

Eva Lasarcyk, Jürgen Trouvain

Institute of Phonetics, Saarland University, Germany
{evaly|trouvain}@coli.uni-saarland.de

ABSTRACT

In this study we present initial efforts to model laughter with an articulatory speech synthesizer. We aimed at imitating a real laugh taken from a spontaneous speech database and created several synthetic versions of it using articulatory synthesis and diphone synthesis. In modeling laughter with articulatory synthesis, we also approximated features like breathing noises that do not normally occur in speech.

Evaluation with respect to the perceived degree of naturalness indicated that the laugh stimuli would pass as "laughs" in an appropriate conversational context. In isolation, though, significant differences could be measured with regard to the degree of variation (durational patterning, fundamental frequency, intensity) within each laugh.

Keywords: Laughter synthesis, articulatory synthesis, synthetic laughter evaluation.

1. INTRODUCTION

Enriching synthetic speech with paralinguistic information including non-verbal vocalizations such as laughter is one of the important challenges in current speech synthesis research. The modeling of laughter has been attempted for concatenative synthesis [4, 12] and formant synthesis [10].

We present an initial study to find out whether articulatory synthesis is a viable alternative. To this end, we analyze the articulation of laughter and create three synthetic laughs on the basis of this analysis. The synthetic laughs differ with respect to degree of variation and with respect to the synthesis method used (see also Sec. 1.2).

The second goal of this study is to investigate if the variation of the details of a laugh (e.g. fundamental frequency, intensity, durational patterning) increases the degree of perceived naturalness of the laugh.

We present a perceptual evaluation that tested, firstly, whether our laugh imitations are "good" enough to pass as a laugh in conversation, and, secondly, to find out whether the rating in naturalness improves when we put more variation into the modeling of a laugh.

1.1. Laughter

A laugh as a whole is very rich in variation and very complex. There are, however, attempts (see e.g. [11] for an overview) to categorize different types of laughter. Bachorowski *et al.* [1] for example introduced three types of human laughs: *song-like*, *snort-like*, and *unvoiced grunt-like*. We will concentrate on the song-like type, "consisting primarily of voiced sounds", including "comparatively stereotyped episodes of multiple vowel-like sounds with evident F_0 modulation …" (p. 1583).

Categorizations have to focus on high level descriptions but authors emphasize at the same time that laughter is *not* a stereotypical sequence of laugh sounds [1, 5]. In [5], Kipper and Todt state that acoustic features like fundamental frequency (F_0), intensity, and tempo (durational pattern) as well as their changing nature "seem to be crucial for the identification and evaluation" of a laugh (p. 256).

Regarding re-synthesized human laughs, Kipper and Todt [5] found that stimuli were rated most positively when they contained varying acoustic parameters (p. 267), which in their case were the durational pattern (rhythm) and the fundamental frequency (pitch).

While a laugh event *itself* can be described, one has to take into account that laughter naturally occurs in a phonetic *context*. The preceding stretch of speech of the laughing person *himself/herself* influences the characteristics of the laugh. It is important, for instance, to match the degree of intensity of the laugh with its phonetic context [12]. Otherwise a laugh would be easily perceived as inappropriate. The phonetic context can also be the utterances of the dialog *partner* where a too intense laugh would be equally inappropriate.

1.2. Synthesis

We used two different synthesis programs to synthesize our laugh samples. One of them was an articulatory speech synthesis system [3], the other one was a diphone synthesis system [8] (see Sec. 3.1 and 3.2). However, the main emphasis was put on the use of the articulatory system. Since the diphone system draws its speech material from prerecorded regular speech (excluding laughs etc.), it obviously cannot be as flexible as a synthesizer that simulates the whole production process. It was mainly used here to delineate the possible advantages (or disadvantages) of the articulatory system.

2. ANALYSIS

2.1. Database

We intended to synthesize a *detailed* laugh and therefore decided to *imitate* natural models of laughter events of spontaneous conversations with overlapping speech. We used a corpus where the two speakers of a dialog were recorded on different audio channels simultaneously [6]. The selected conversation by two male speakers contained 13 situations with one or more laughs and we focused on the *song-like* type of laugh.

2.2. Features of the laugh

Descriptions in [1] concentrate primarily on what we will be calling the *main part* of the laugh (see below). While their definition is plausible for some research questions, we wish to extend the definition of a laugh to include breathing and pausing. Audible breathing can often be observed, framing the *main part* and *pause* of a laugh in the corpus. Since the articulatory synthesizer should be able to generate breath noises, we take this feature into account.

The following structure is thus proposed for the laughs analyzed and imitated in this study:

- an *onset* (an audible forced exhalation [7]),
- a *main part* with laugh syllables, each containing a voiced and an unvoiced portion,
- a *pause*, and
- the *offset*, consisting of at least one audible deep inhalation.

To see a human laugh labeled according to these four phases please refer to image file 1 (top).

In order to re-synthesize the laugh, the following items were specified:

- duration of the *onset*, each laugh syllable in the *main part*, the *pause*, and the *offset*,
- intensity contour of the whole laugh,
- fundamental frequency contour of the laugh,
- vowel quality of the voiced parts.

2.3. Overall results of the analysis

Image file 2 (a) shows a colored screenshot (using the software in [9]) of an oscillogram and a spectrogram of a human laugh from the corpus used here (cf. audio file 1). F_0 and intensity contours are visible in the colored spectrogram (blue and yellow lines.)

The temporal succession of elements can be seen as labels in image file 1: The first element of the laugh (*onset*) is an audible exhalation. This is followed in a *main part* by several laugh syllables of decreasing overall intensity and increasing overall length. Within a laugh syllable, an energy-rich portion (voiced) is followed by a breathy portion (unvoiced), later on with faint sounds in between. The *main part* is followed by a *pause*. The last element of the laugh (*offset*) is a forced inhalation to compensate for the low lung volume.

2.4. Some physiological details

The following physiological and articulatory aspects are important for the control of the articulatory synthesizer.

2.4.1. Subglottal pressure

Luschei et al. [7] state that "laughter generally takes place when the lung volume is low" (p. 442). Nevertheless, the tracheal pressure during laughs can reach peaks of around 1.8 to 3.0 kPa (p. 446), which is higher than the level typical of speech.

2.4.2. Vowel quality

The vowel quality of the voiced portion of a laugh syllable must be defined. Bickley and Hunnicutt [2] found that the formant patterns "do not appear to correspond to a standard … vowel" (p. 929) of the laughers' mother tongue but do fall into the normal range of speakers' formant values. Bacharowski *et al.* [1] found that their recorded laughs generally contained "central, unarticulated sounds" (p. 1594).

3. SYNTHESIS

To imitate the human laugh, we used two different synthesis systems both of which have their merits.

3.1. Articulatory synthesis

One system was the articulatory synthesis system described in [3]. The speech output is generated from a gestural score (containing several *tiers*, as can be seen in image file 1) via an aerodynamic-acoustic simulation of airflow through a 3D model of the vocal tract. This allows for a high degree of freedom and control over a number of parameters including subglottal pressure, vocal tract shapes, and different glottal settings. With this type of synthesis it is thus also possible, in principle, to create breathing noise and freely approximate virtually any vowel quality needed.

3.2. Diphone synthesis

The second system was the diphone system MARY [8]. Speech is generated by choosing, manipulating and concatenating appropriate units from a corpus of prerecorded and segmented natural speech. The output is thus based on natural human speech. Since the set of sounds is limited by the corpus, it is not possible to imitate the breathing portions of the laugh, and for the laugh syllables only the predefined phones are available.

3.3. Imitating laughter in different versions

In the following section, we describe the generation of the three different imitations of the human laugh (*version H*) shown in image file 2a.

3.3.1. Version V

Of all three synthetic versions, *version V* (image file 2b, audio file 2) contained the highest degree of variation within the laugh in terms of durational patterning, intensity and F_0 contours. The duration of each of the phases and of each laugh syllable within the *main part* was copied from the human laugh sample. Intensity and F_0 movements (yellow and blue lines in the image) were also modeled in a way to match the human ones as closely as possible.

In each laugh syllable in the *main part*, voiced and unvoiced portions alternate. To reflect this basic pattern of vocalization, glottal gestures were placed alternately on the glottal gesture tier in the gestural score (see bottom of image file 1). An "open" gesture corresponds to the unvoiced portion of a laugh syllable, a "close" gesture to the voiced portion ("laugh vowel" [11]). The duration of each gesture was copied from the durational patterning of the human laugh.

To get the appropriate vowel quality in the *main part*, a vowel gesture was placed on the vocalic tier so that when the glottis is ready for phonation, a laugh vowel would be articulated. We approximated the speaker in our sample laugh by using an [ɛ] on the vocalic tier.

In order to model the different levels of intensity within the *main part*, we varied the degree of lung pressure by using different gestures on the pulmonic pressure tier (bottom tier).

The overall (long-term) F_0 contour was modeled with appropriate gestures on the F_0 phrase tier. F_0 accent gestures were used to imitate the (short-term) fundamental frequency contour within one laugh syllable.

Since the kind of laugh imitated here also contains two breathing phases (*onset* and *offset*), we put gestures of generally high lung pressure on the pulmonic tier and gestures of a widely abducted position of the vocal folds ("open") on the glottal tier. The result was, however, a long way from the original level of intensity. Thus, an additional source of friction was introduced on the consonantal tier ("E:_Pharynx"). This implies a constriction area in the pharynx, and was motivated by introspection, analogous to constrictions in grunt-like laughs [1]. The result was a clearly audible friction noise.

3.3.2. Version S

The second imitation created with the articulatory synthesizer was *version S* (cf. audio file 3, image file 2c). It contained *less* variation in durational patterning, intensity, and fundamental frequency in the *main part*.

The gestural score for this version was constructed by taking *version V* and deleting all the (variation-rich) *main part* gestures except for the gestures of the first laugh syllable. The gap was then filled by repeating the block of gestures for the first laugh syllable until this laugh imitation contained the same number of laugh syllables as the human one and *version V*. *Version S* was thus a more stereotypical imitation than *version V*.

3.3.3. Version D

Due to the inherent phone set restrictions of a diphone synthesis system, the diphone *version D* (audio file 4, image file 2d) was generated without the breathing phases (*onset* and *offset*). As a consequence, the phase containing the pause would become obsolete since no signal followed. The

main part of *version D* was produced by alternating the phones [ε] and [h], which seemed to resemble best the unvoiced and voiced portions of each laugh syllable. The durational pattern was, as in *version V*, adopted from the human laugh.

The fundamental frequency contour was approximated by specifying a target frequency value for each of the [ε] and [h] segments. We did not have explicit control over intensity values.

4. PERCEPTUAL EVALUATION

We carried out two perception experiments to get ratings of how natural the laughs would be perceived. In the first experiment, the laughs were integrated in a dialog, whereas in the second experiment, they were presented in isolation.

4.1. Stimuli

For the first experiment, the aim was to keep the verbal interaction presented as natural as possible by placing the synthesized laugh versions at exactly the same location as the original (human) one. Audio file 5 contains the dialog in its original version, 6 to 8 with laugh *versions V, S,* and *D,* respectively. The dialog structure of the stimuli was always identical: Person 1 speaks and laughs, directly afterwards, about his own statement; person 2 joins in. In one stimulus, this laugh of person 2 is human (original, version *H*), the other three each contain one of the synthetic laughs.

For the second experiment, each of these four laughs (one human, three synthetic) was prepared to be presented in isolation by cutting it out of the conversational context. The aim of presenting them in this isolated way was to allow for a more direct focus on the laugh itself in order to asses its intrinsic naturalness. The human laugh (audio file 1) obviously contained the highest degree of variation, *version V* a mid-high degree of variation, and *versions S* and *D* contained less variation (regarding durational patterns, intensity and fundamental frequency).

4.2. Experimental setup and participants

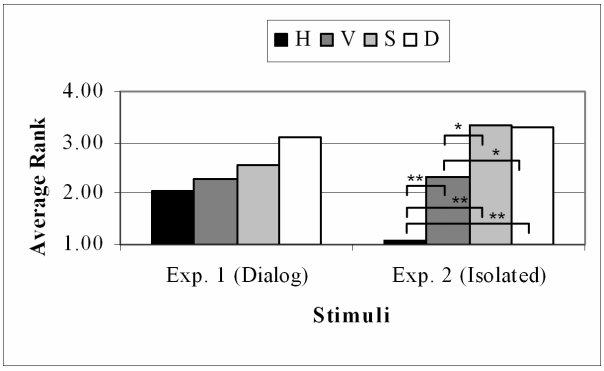
The experiments were conducted together, one immediately after the other. All participants (14 in total, 8 female, 6 male, with an average age of 25 years) participated in both sessions. The audio material was presented to each person individually via loudspeakers in a separate randomized order for each participant to minimize order effects. The participants were asked to rate each stimulus with respect to naturalness on a scale of 1 to 4: 1 "natural", 2 "less natural", 3 "rather unnatural", and 4 "unnatural". Thus, in experiment 1, they were asked to give their overall impression of how natural they found the dialog in total. In experiment 2, they were asked to rate the naturalness of the laugh stimulus by itself.

For both experiments, we calculated the average ranks of each stimulus (dialog or laugh). A non-parametric Friedman test (significance threshold 5 %) was applied to ascertain significant effects of laugh type within an experiment.

For experiment 1, the null hypothesis was: There is no dependency between the rating of a dialog and the laugh stimulus placed in the dialog. The alternative hypothesis was: The rating of the dialogs depends on which laugh stimulus is placed into them.

For experiment 2, the null hypothesis was: There is no dependency between the rating of an isolated laugh and its degree of internal variation. The alternative hypothesis was: The rating of an isolated laugh depends on how rich its internal variation is.

Fig. 1 Average ranks regarding naturalness in experiments 1 and 2. Bars between pairs mark significant differences of $p < 0.0083$ (*) and $p < 0.001$ (**). Properties of the stimuli *H, V, S,* and *D* are explained in Sec. 3.3.



4.3. Results

Fig. 1 shows the average ranks of the ratings for experiment 1 (left) and 2 (right).

The dialog stimuli of experiment 1 were ranked in the following order: *H* (average rank of 2.07), *V* (2.29), *S* (2.54), and *D* (3.11). It has to be added that the ratings for this experiment did *not* differ significantly.

For experiment 2, the order of the stimuli was similar, only the last two were reversed: *H* (1.07),

V (2.32), *D* (3.29), and *S* (3.32). In this experiment, the ratings differed significantly. Thus, we conducted post-hoc pair-wise comparison tests (Wilcoxon) to determine which versions differed significantly from one another. The 5 % significance threshold was corrected to 0.83 % since we had 6 pairs to compare.

We found a significant difference between all the pairs except between stimuli *S* and *D*. *H* was ranked as significantly more natural than *V*, *S*, and *D* ($p < 0.001$). *V* was ranked as significantly more natural than *S* ($p = 0.002$) and *D* ($p = 0.008$). The more natural rating of *D* with respect to *S* was *not* significant ($p = 0.688$).

5. DISCUSSION

5.1. Experiment 1

The outcome of experiment 1 might indicate that all synthetic laughs are "good enough" to pass as laughter in the dialog.

This is especially noteworthy with respect to the laugh *version D*: It was created with a diphone synthesis system that can assemble a laugh only from regular speech sounds. This may in a way support the indication Bickley and Hunnicutt found in their study [2] that "in some ways laughter is speech-like" (p. 930) since they found *similar* measurements of the temporal and spectral characteristics of their *laughs* to what they found in *speech*. It may also indicate that the natural (human) origin of the diphones to a certain extent counterbalances the purely synthetic voice of the articulatory system. Sounding more natural *per se* may be advantageous; another issue is the degree of flexibility (discussed below in Sec. 5.2).

It can be argued, though, that the context chosen here was *masking* the target laugh too much, and that the major part of the dialog was made up of unprocessed natural speech/laughing. This can be seen as an "advantage", yielding relatively high values of naturalness. Nevertheless, it was a real-life context, and joint laughter of two speakers is presumably not uncommon [12]. Still the question arises: What other context would be better suited to the test?

Another point of discussion is the fact that, in experiment 1, the participants were asked to rate the naturalness of the dialog *as a whole*. Our initial intention had been to compare a laugh *within a dialog* with a laugh *in isolation*. In order to do this, it might have been possible to address the laugh item in the dialog directly, when giving the instructions, and in this way create a bias in the expectation of the listener. However, we did not want to influence the participants before they heard the dialog by saying that it contained laughter. Thus, we could not compare the ratings directly with those of experiment 2.

5.2. Experiment 2

The results of experiment 2 indicate, firstly, that all synthetic versions are perceived as much less natural than the natural version. This can be expected, since natural speech introduces an extremely high standard and laughs in particular can be very complex. Furthermore, the synthetic stimuli created here were an initial approach to modeling laughter.

Secondly, while *all* synthetic stimuli in our experiments seemed "good enough" to pass as laughter in speaker-overlapping context, presenting them in isolation brought to light that there are differences in perceived naturalness with regard to the *variation* within a laugh. The significantly better (i.e. more natural) ranking of *version V* suggests that, in principle, it should be possible to improve perceived naturalness by putting more details and variations into a laugh stimulus. This result may be seen as confirmation of previous findings; see e.g. the overview and study in [5] which concludes that variation within a laugh is important for its evaluation.

It can be argued, though, that the *version D* and *version S* laughs sound rather simple and in consequence, the better rating of *version V* should not come as a surprise. The stimuli *D* and *S* were meant to be reasonable initial imitations of the human laugh, though with less variation than *V*. Some features were impossible to model in the diphone synthesis system, such as the breathing noise, the selection of the "laugh vowel", or the lack of intensity control. Other features were deliberately generated in a less varied way (such as the durational pattern, fundamental frequency, and intensity in *version S*). Maybe a more fine-grained scale of variation could be designed and implemented in laugh stimuli synthesis in the future.

Another dimension in the discussion is whether articulatory synthesis provides any advantages when imitating laughter. In general, synthesizing laughter "from scratch" in an articulatorily transparent way seems quite promising, the reason being that with the different gestures one could

model the articulation processes quite directly – we have to note, though, that the gestural solutions used here do not necessarily mirror correctly what humans do when producing laughter. The results of experiment 1 might only indicate that this is *one* way of doing it.

Apart from the advantage of modeling gestures directly, we also noted limitations to the current articulatory approach. The first is of a more technical nature. E.g. the current limit of 1 kPa to the pulmonic pressure is appropriate for speech but seemingly not high enough for laughter. In this case we compensated by introducing the *ad hoc* constriction in the pharynx in order to achieve the desired level of friction noise. This choice might not reflect accurately what really happens during laughter.

The second kind of limitation stems from our limited knowledge of some aspects of laughing. We need to know exactly what sort of excitation there is at the glottis. When modeling singing, we add tremolo to the voice; what could be the adequate or necessary additions to the regular source signal when modeling laughter?

6. CONCLUSIONS

Imitating human laughter in conversation proves to be a challenging field when it comes down to modeling the articulatory aspects of laughter, not all of which are known in full detail yet. The general approach seems promising and the perceptual tests conducted suggest that the articulatory synthesizer used for our stimuli is indeed capable of producing purely synthetic laugh-like sounds with varying degrees of variation.

It therefore presents a viable alternative to other forms of parametric laughter synthesis like formant synthesis [10]. In contrast to concatenative synthesis, more room for improvement and fine-tuning exists.

In concatenative systems, the continuum of possible variation is limited. A *regular* diphone synthesis system, for example, relies on speech sounds only. Thus, only (stylized) "haha" laughs are possible, restricting the set of possible variations to fundamental frequency, duration, and the phone choice of the laugh vowels.

In a further approach, whole prerecorded laughs are inserted into concatenative speech, either in combination with diphone speech [12] or as autonomous units in unit-selection synthesis [4]. However, the laughs must either be selected according to yet unknown criteria or they must be manipulated again in ways with unclear phonetic results for the listener. It is easy to sound ridiculous with the wrong laugh.

Further work could include the generation of laugh stimuli with articulatory synthesis that allow for more detailed testing of different features, varied with respect to intensity, fundamental frequency, breathing noise, friction sources, one or more laugh vowels etc. Several goals could be pursued: The set of articulatory gestures that work best for imitating particular laughs could be investigated, or articulatory synthesis could be used to build systematically varying laugh stimuli to test the impact that particular features have on the listener.

Another aspect associated with laughter is the question of *speech laughs*, i.e., where laughing occurs simultaneously with speech. It would be a highly challenging task to undertake with the articulatory synthesizer.

7. REFERENCES

[1] Bachorowski, J.-A., Smoski, M.J., Owren, M.J. 2001. The acoustic features of human laughter. *Journal of the Acoustical Society of America* 110, 1581-1597.

[2] Bickley, C., Hunnicutt, S. 1992. Acoustic analysis of laughter. Proc. *2nd International Conference on Spoken Language Processing*, Banff (2), 927-930.

[3] Birkholz, P. 2006. *3D-Artikulatorische Sprachsynthese.* Logos, Berlin. (PhD thesis).

[4] Campbell, N. 2006. Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Audio, Speech, and Language Processing* 14, 1171-1178.

[5] Kipper, S., Todt, D. 2003. The role of rhythm and pitch in the evaluation of human laughter. *Journal of Nonverbal Behavior* 27, 255-272.

[6] Kohler, K. J., Peters, B., Scheffers, M. 2006. *The Kiel Corpus of Spontaneous Speech Vol. IV. German: Video Task Scenario (Kiel-DVD #1).* http://www.ipds.uni-kiel.de/kjk/pub_exx/kk2006_6/InfoDVD1.pdf visited 29-Mar-07

[7] Luschei, E.S., Ramig, L.O., Finnegan, E.M., Baker, K.K., Smith, M.E. 2006. Patterns of laryngeal electromyography and the activity of the respiratory system during spontaneous laughter. *J Neurophysiology* 96, 442-450.

[8] *MARY* Text-to-Speech Synthesis System. http://mary.dfki.de/online-demos/speech_synthesis visited 28-Mar-07

[9] *Praat*: Doing Phonetics by Computer. *www.praat.org* (version: 4.5.14) visited 05-Feb-2007

[10] Sundaram, S., Narayanan, S. 2007. Automatic acoustic synthesis of human-like laughter. *Journal of the Acoustical Society of America* 121 (1), 527-535.

[11] Trouvain, J. 2003. Segmenting phonetic units in laughter. *Proc. 15th. International Conference of the Phonetic Sciences*, Barcelona, 2793-2796.

[12] Trouvain, J., Schröder, M. 2004. How (not) to add laughter to synthetic speech. Proc. of the *Workshop on Affective Dialogue Systems*, Kloster Irsee, 229-232.

Evaluating automatic laughter segmentation in meetings using acoustic and acoustic-phonetic features

Khiet P. Truong and David A. van Leeuwen

TNO Human Factors
P.O. Box 23, 3769 ZG, Soesterberg, The Netherlands
{khiet.truong, david.vanleeuwen}@tno.nl

ABSTRACT

In this study, we investigated automatic laughter segmentation in meetings. We first performed laughter-speech discrimination experiments with traditional spectral features and subsequently used acoustic-phonetic features. In segmentation, we used Gaussian Mixture Models that were trained with spectral features. For the evaluation of the laughter segmentation we used time-weighted Detection Error Tradeoff curves. The results show that the acoustic-phonetic features perform relatively well given their sparseness. For segmentation, we believe that incorporating phonetic knowledge could lead to improvement. We will discuss possibilities for improvement of our automatic laughter detector.

Keywords: laughter detection, laughter

1. INTRODUCTION

Since laughter can be an important cue for identifying interesting discourse events or emotional user-states, laughter has gained interests from researchers from multidisciplinary research areas. Although there seems to be no *unique* relation between laughter and emotions [12, 11], we all agree that laughter is a highly communicative and social event in human-human communication that can elicit emotional reactions. Further, we have learned that it is a highly variable acoustic signal [2]. We can chuckle, giggle or make snort-like laughter sounds that may sound differently for each person. Sometimes, people can even identify someone just by hearing their laughter. Due to its highly variable acoustic properties, laughter is expected to be difficult to model and detect automatically.

In this study, we will focus on laughter recognition in speech in meetings. Previous studies [6, 13] have reported relatively high classification rates, but these were obtained with either given pre-segmented segments or with a sliding n-second window. In our study, we tried to localize spontaneous laughter in meetings more accuractely on a frame basis. We did not make distinctions between different types of laughter, but we rather tried to build a generic laughter model. Our goal is to automatically detect laughter events for the development of affective systems. Laughter event recognition implies automatically positioning the start and end time of laughter. One could use an automatic speech recognizer (ASR) to recognize laughter which segments laughter as a by-product. However, since the aim of an automatic speech recognizer is to recognize speech, it is not specifically tuned for detection of non-verbal speech elements such as laughter. Further, an ASR system employing a full-blown transcription may be a bit computationally inefficient for the detection of laughter events. Therefore, we rather built a relatively simple detector based on a small number of acoustic models. We started with laughter-speech discrimination (which was performed on pre-segmented homogeneous trials), and subsequently, performed laughter segmentation in meetings. After inspection of some errors of the laughter segmentation in meetings, we believe that incorporating phonetic knowledge could improve performance.

In the following sections we desribe the material used in this study (Section 2), our methods (Section 3) and we explain how we evaluated our results (Section 4). Subsequently, we show our results (Section 5) and discuss how we can improve laughter segmentation (Section 6).

2. DATABASE

We used spontaneous meetings from the ICSI Meeting Recorder Corpus [8] to train and test our laughter detector (Table 1). The corpus consists of 75 recorded meetings with an average of 6 participants per meeting and a total of 53 unique speakers. We used the close-talk recordings of each participant. The first 26 ICSI 'Bmr' ('Bmr' is a naming convention of the type of meeting at ICSI) meetings were used for training and the last 3 ICSI 'Bmr' meetings (10 unique speakers, 2 female and 8 male) were used for testing. Some speakers in the training set were also present in the test set. Note that the manually produced laughter annotations were not always precise, e.g., onset and offset of laughter were not always marked.

Table 1: *Amount of data used in our analyses (duration, numbers of segments in brackets).*

	Training 26 Bmr meetings	Testing 3 Bmr meetings
Speech	81 min (2422)	10 min (300)
Laughter	83 min (2680)	10 min (279)

For training and testing, we used only audible laughter events (relatively clearly as perceived by the first author). The segments consisted of solely audible laughter which means that so-called "speech-laughs" or "smiled speech" was not investigated.

3. METHOD

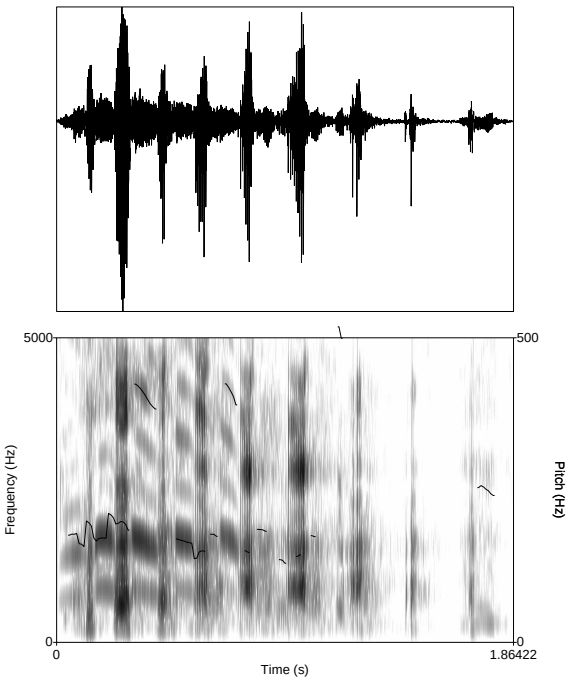
3.1. Acoustic modeling

3.1.1. Laughter-speech discrimination

For laughter-speech discrimination, we used cepstral and acoustic-phonetic features. Firstly, Gaussian Mixture Models (GMMs) were trained with Perceptual Linear Prediction Coding (PLP) features [5]. Twelve PLP coefficients and one log energy component, and their 13 first order derivatives (measured over five consecutive frames) were extracted each 16 ms over a window with a length of 32 ms. A 'soft detector' score is obtained by determining the log likelihood ratio of the data given the laughter and speech GMMs respectively.

Secondly, we used utterance-based acoustic-phonetic features that were measured over the whole utterance, such as mean log F_0, standard deviation of log F_0, range of log F_0, the mean slope of F_0, the slope of the Long-Term Average Spectrum (LTAS) and the fraction of unvoiced frames (some of these features have proven to be discriminative [13]). These features were all extracted with PRAAT [4]. Linear Discriminant Analysis (LDA) was used as a discrimination method which has as advantage that we can obtain information about the contribution of each feature to the discriminative power by examining the standardized discriminant coefficients which can be interpreted as feature weights. The posterior probabilities of the LDA classification were used as 'soft detector' scores. Statistics of F_0 were chosen because some studies have reported significant F_0 differences between laughter and speech [2] (although contradictory results have been reported [3]). A level of 'effort' can be measured by the slope of the LTAS: the less negative the slope is, the more vocal effort is expected [9]. And the fraction of unvoiced frames was chosen since due to the characteristic alternating voicing/unvoicing pattern which is often present in laughter, it is expected that the percentage of unvoiced frames is larger in laughter than in speech (which was suggested by [3]), see Fig. 1. Note that measures of F_0 can only be measured in the vocalized parts of laughter. A disadvantage of such features is that they cannot easily be used for a segmentation problem because these features describe relatively slow-varying patterns in speech that require a larger time-scale for feature extraction (e.g., an utterance). In segmentation, a higher resolution of extracted features (e.g., frame-based) is needed because accurate localization of boundaries of events is important.

Figure 1: *Example of laughter with typical voiced/unvoiced alternating pattern, showing a waveform (top) and a spectrogram (bottom).*



3.1.2. Laughter segmentation

For laughter segmentation, i.e., localizing laughter in meetings, we used PLP features and trained three GMMs: laughter, speech and silence. Silence was added because we encountered much silence in the meetings, and we needed a way to deal with it. In order to determine the segmentation of the acoustic signal into segments representing the N defined classes (in our case $N = 3$) we used a very simple Viterbi decoder [10]. In an N-state parallel topology the decoder finds the maximum likelihood state sequence. We used the state sequence as the segmentation result. We controlled the number of state transitions, or the segment boundaries, by using a small state transition probability. The state transi-

tion probability a_{ij} from state i to state $j \neq i$ were estimated on the basis of the average duration of the segments i and the number of segments j following i in the training data. The self probabilities a_{ii} were chosen so that $\sum_j a_{ij} = 1$. After the segmentation into segments $\{s_i\}$, $i = 1, \ldots N_s$, we calculated the average log-likelihoods L_{im} over each segment i for each of the models m. We defined a log-likelihood-ratio as $L_{laugh} - \max(L_{speech}, L_{silence})$. These log-likelihood-ratios determine final class-membership.

4. EVALUATION METRIC

For laughter-speech discrimination, we used the Equal Error Rate (EER) as a single performance measure, adopted from the detection framework. In laughter-speech discrimination, we can identify two types of errors: a *false alarm*, i.e., a speech segment is falsely detected as laughter, and a *miss*, i.e., a laughter segment is incorrectly detected as speech. The EER is defined as the error rate where the false alarm rate is equal to the miss rate.

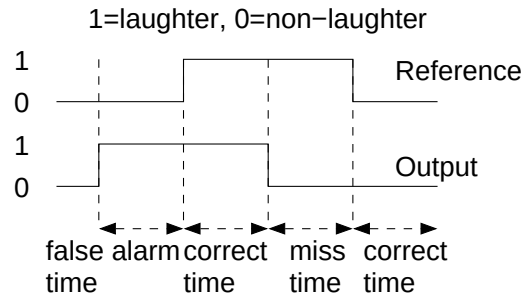
The evaluation of the automatic laughter *segmentation* was not so straightforward. One of the reasons to define log-likelihood ratios for the segments found by the detector, is to be able to compare the current results based on segmentation to other results that were obtained with given pre-segmented segments and that were evaluated with a trial-based DET analysis (Detection Error Tradeoff [7]). In this analysis we could analyze a detector in terms of DET plots and post-evaluation measures such as Equal Error Rate and minimum decision costs. In order to make comparison possible we extended the concept of the trial-based DET analysis to a time-weighted DET analysis for two-class decoding [14]. The basic idea is (see Fig. 2) that each segment in the hypothesis segmentation may have sub-segments that are either

- correctly classified (hits and correct rejects)
- missed, i.e., classified as speech (or other), while the reference says laughter
- false alarm, i.e., classified as laughter, while the reference says speech (or other)

We can now form tuples (λ_i, T_i^e) where T_i^e is the duration of the sub-segment of segment i and e is the evaluation over that sub-segment, either 'correct', 'missed' or 'false alarm'. These tuples can now be used in an analysis very similar to the DET analysis. Define θ as the treshold determining the operating point in the DET plot. Then the false alarm probability is estimated from the set $\mathcal{T}_\theta$ of all tuples for which $\lambda_i > \theta$

$$(1) \quad p_{\text{FA}} = \frac{1}{T_{\text{non}}} \sum_{i \in \mathcal{T}_\theta} T_i^{\text{FA}}$$

Figure 2: *Definitions of correct classifications and erroneous classifications in time.*



and similarly the miss probability can be estimated as

$$(2) \quad p_{\text{miss}} = \frac{1}{T_{\text{tar}}} \sum_{i \notin \mathcal{T}_\theta} T_i^{\text{miss}}$$

Here T_{tar} and T_{non} indicate the total time of target class (laughter) and non-target class (e.g., speech) in the reference segmentation.

5. RESULTS

We tested laughter-speech discrimination and laughter segmentation on a total of 27 individual channels of the close-talk recordings taken from three ICSI 'Bmr' meetings. For laughter-speech discrimination we tested with pre-segmented laughter and speech segments, while for laughter segmentation, full-length channels of whole meetings were applied. The scores (log-likelihood ratios or posterior probabilities) obtained in these audio channels were pooled together to obtain EERs, Table 2. In order to enable better comparison between the laughter-speech discrimination and the laughter segmentation results, we have also performed a segmentation experiment in which we concatenated the laughter and speech segments (used in the discrimination task) randomly to each other and subsequently performed laughter segmentation on this chain of laughter-speech segments. Thus the difference in performance in Fig. 3 is mainly caused by the presence of other sounds, such as silence, in meetings. A disadvantage of the time-weighted DET curve (used for laughter-segmentation) is that it does not take into account the absolute number of times there was an error.

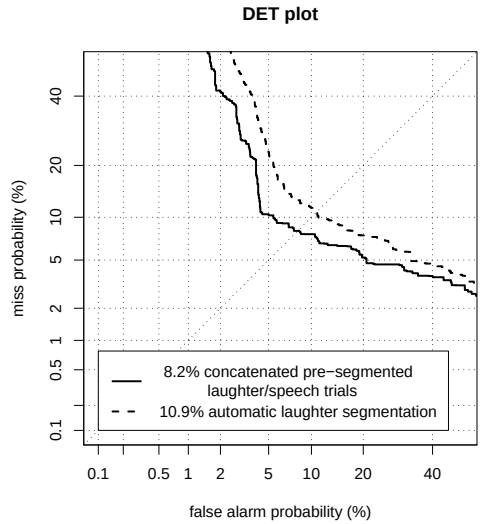
Many of the errors in laughter segmentation were introduced by sounds like, e.g., breaths, coughs, background noises or crosstalk (softer speech from other participants). It seems that, especially, unvoiced units in laughter can be confused with these type of sounds (and vice versa).

The LDA analysis with the PRAAT- features in the laughter-speech discrimination indicated that

Table 2: *EERs of laughter-speech discrimination and laughter segmentation (tested on 3 ICSI Bmr meetings). The lower the EERs, the better the performance.*

Discrimination Pre-segmented		Segmentation Concatenated laughter/speech	Whole meetings
GMM PLP	LDA PRAAT	GMM PLP	GMM PLP
0.060	0.118	0.082	0.109

Figure 3: *Time-weighted DET curves of laughter segmentation, tested on 3 ICSI Bmr meetings.*



mean log F_0 and the fraction of unvoiced frames had the highest weights, which means that these two features contributed the most discriminative power to the model. The LDA model in combination with these features seem to perform relatively well, given the small number of features used.

6. DISCUSSION AND CONCLUSIONS

We believe that the performance of the laughter segmenter can be improved by incorporating phonetic knowledge into the models. In a previous study [13], a fusion between spectral and acoustic-phonetic features showed significant improvement in laughter-speech discrimination. However, acoustic-phonetic features are usually measured over a longer timescale which makes it difficult to use these for segmentation. Currently, we are modeling laughter as a whole with GMMs that are basically one-state Hidden Markov Models (HMMs). The results of the LDA analysis indicate that we could employ phonetic information about the voiced (where we can measure F_0) and unvoiced parts of laughter (the fraction of unvoiced frames appeared to be discriminative). We could use HMMs to model subcomponents of laughter which are based on phonetic units, e.g., a VU (voiced-unvoiced) syllable could be such a phonetic unit. With HMMs, we can then better model the time-varying patterns of laughter, such as the characteristic repeating /haha/ pattern by the HMM state topology and state transition probabilities. However, for this purpose, a large database containing different laughter sounds which are annotated on different phonetic levels is needed. In addition, our laughter segmentation model may be too generic. We could build more specific laughter models for, e.g., voiced laughter, which appears to be perceived as 'more positive' by listeners [1]. Further, we have used a time-weighted DET analysis which has as an important advantage that it has a DET-like behavior so that comparisons between other studies that use DET analyses are easier to make. Disadvangtages are that it does not take into account the number of times that a detector has made an error, and our time-weighted evaluation could have been too strict (it is not clear what exactly defines the beginning and end of laughter).

We are currently implementing an online laughter detector which will be used in an interactive affective application. Additional challenges arose during the development of our online laughter detector, such as how to perform online normalization. In the future, we intend to improve our laughter detector by employing more phonetic properties of laughter.

ACKNOWLEDGEMENTS

This work was supported by a Dutch Bsik project: MultimediaN.

7. REFERENCES

[1] Bachorowski, J.-A., Owren, M. 2001. Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect. *Psychological Science* 12, 252–257.

[2] Bachorowski, J.-A., Smoski, M., Owren, M. 2001. The acoustic features of human laughter. *J.Acoust.Soc.Am.* 110, 1581–1597.

[3] Bickley, C., Hunnicutt, S. 1992. Acoustic analysis of laughter. *Proc. ICSLP* 927–930.

[4] Boersma, P. 2001. Praat: system for doing phonetics by computer. *Glot International*.

[5] Hermansky, H. 1990. Perceptual linear predictive (PLP) analysis of speech. *J.Acoust.Soc.Amer.* 87, 1738–1752.

[6] Kennedy, L., Ellis, D. 2004. Laughter detection in meetings. *NIST ICASSP 2004 Meeting Recognition Workshop* 118–121.

[7] Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M. 1997. The DET curve in as-

sessment of detection task performance. *Proc. Eurospeech* 1895–1898.
[8] Morgan, N., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Janin, A., Pfau, T., Shriberg, E., Stolcke, A. 2001. The meeting project at ICSI. *Proc. Human Language Technologies Conference* 1–7.
[9] Pittam, J., Gallois, C., Callan, V. 1990. The long-term spectrum and perceived emotion. *Speech Communication* 9, 177–187.
[10] Rabiner, L., Juang, B. 1986. An introduction to Hidden Markov Models. *IEEE ASSP Magazine* 3, 4–16.
[11] Russell, J., Bachorowski, J., Fernandez-Dols, J. 2003. Facial and vocal expressions of emotion. *Annu.Rev.Psychology* 54, 329–349.
[12] Schröder, M. 2003. Experimental study of affect bursts. *Speech Communication* 40, 99–116.
[13] Truong, K., Van Leeuwen, D. 2005. Automatic detection of laughter. *Proc. Interspeech* 485–488.
[14] Van Leeuwen, D., Huijbregts, M. 2006. The AMI speaker diarization system for NIST RT06s meeting data. *Proc. MLMI* 371–384.

On the Correlation between Perceptual and Contextual Aspects of Laughter in Meetings

Kornel Laskowski and Susanne Burger

interACT, Carnegie Mellon University, Pittsburgh PA, USA
kornel|sburger@cs.cmu.edu

ABSTRACT

We have analyzed over 13000 bouts of laughter, in over 65 hours of unscripted, naturally occurring multiparty meetings, to identify discriminative contexts of voiced and unvoiced laughter. Our results show that, in meetings, laughter is quite frequent, accounting for almost 10% of all vocal activity effort by time. Approximately a third of all laughter is unvoiced, but meeting participants vary extensively in how often they employ voicing during laughter. In spite of this variability, laughter appears to exhibit robust temporal characteristics. Voiced laughs are on average longer than unvoiced laughs, and appear to correlate with temporally adjacent voiced laughter from other participants, as well as with speech from the laugher. Unvoiced laughter appears to occur independently of vocal activity from other participants.

1. INTRODUCTION

In recent years, the availability of large multiparty corpora of naturally occurring meetings [2] [7] [3] has focused attention on previously little-explored, natural human-human interaction behaviors [17]. A non-verbal phenomenon belonging to this class is laughter, which has been hypothesized as a means of affecting interlocutors, as well as a signal of various human emotions [14].

In our previous work, we produced an annotation of perceived emotional valence in speakers in the ISL Meeting Corpus [10]. We showed that instances of isolated laughter were strongly predictive of positive valence, as perceived in participants by external observers who had not participated in the meetings. In a subsequent multi-site evaluation of automatic emotional valence classification within the CHIL project [21], we found that transcribed laughter is in general much more indicative of perceived positive valence than any other grouping of spectral, prosodic, contextual, or lexical features. Three-way classification of speaker contributions into negative, neutral and positive valence classes (with neutral valence accounting for 80% of the contributions), using the presence of transcribed laughter as the only feature, resulted in an accuracy of 91.2%. The combination of other features led to an accuracy of only 87% (similar results were produced on this data by [12]). A combination of all features, including the presence of transcribed laughter, produced an accuracy of 91.4%, only marginally better than transcribed laughter alone.

Although these results show that the presence of laughter, as detected by human annotators, was the single most useful feature for automatic valence classification, laughter and positive valence are not completely correlated in the ISL Meeting Corpus. We are ultimately interested in the ability to determine, automatically, whether a particular laugh conveys information about the laughter's valence to an outside observer. The current work is a preliminary step in that effort, in which we characterize laughter along two separate dimensions. First, we determine whether each laugh is voiced or unvoiced. Previous work with this distinction in other domains has shown that voiced laughter may be used strategically in conversation [14].

Second, we attempt to characterize the temporal context of voiced and unvoiced laughter within the multiparticipant vocal activity on-off pattern of a conversation. In the current work, we are interested exclusively in *text-independent* context, which allows us to ignore specific lexical and/or syntactic phenomena having bearing on the occurrence of laughter. The study of laughter in sequence with spontaneous speech has been treated by conversation analysis [8]; the latter has offered solutions for both transcribing and investigating multiparticipant laughter [9] [4], but it has not produced quantitative descriptions or means of obtaining them. Laughter has also been shown to evoke laughing in listeners [15], in this way differing from speech. In particular, laughers do not take turns laughing in the same way that speakers take turns speaking. Vocal activity context therefore appears to provide important cues as to whether ongoing vocal activity is laughter or speech [11]. In the current work, we attempt to determine whether context also disambiguates between voiced and unvoiced laughter.

2. DATA

To study the pragmatics of laughter, we use the relatively large ICSI Meeting Corpus [7]. This corpus consists of 75 unscripted, naturally occurring meetings, amounting to over 71 hours of recording time. Each meeting contains between 3 and 9 participants wearing individual head-mounted microphones, drawn from a pool of 53 unique speakers (13 female, 40 male); several meetings also contain additional participants without microphones.

In this section, we describe the process we followed to produce, for each meeting and for each participant: (1) a *talk spurt* segmentation; (2) a voiced *laugh bout* segmentation; and (3) an unvoiced laugh bout segmentation. A talk spurt is defined [18] as a contiguous interval of speech delineated by non-speech of at least 500 ms in duration; laugh bouts, as used here, were defined in [1].

We note that each meeting recording contains a ritualized interval of read speech, a subtask referred to as Digits, which we have analyzed but excluded from the final segmentations. The temporal distribution of vocal activity in these intervals is markedly different from that in natural conversation. Excluding them limits the total meeting time to 66.3 hours.

2.1. Talk Spurt Segmentation

Talk spurt segmentation for the meetings in the ICSI corpus was produced using word-level forced alignment information, available in a corpus of auxiliary annotations known as in the ICSI Dialog Act Corpus [19]. While 500 ms was used as the minimum inter-spurt duration in [18], we use a 300 ms threshold. This value has recently been adopted for the purposes of building speech activity detection references in the NIST Rich Transcription Meeting Recognition evaluations.

2.2. Selection of Transcribed Laughter Instances

Laughter is transcribed in the ICSI Meeting Corpus orthographic transcriptions in two ways. First, discrete events are annotated as `VocalSound` instances, and appear interspersed among lexical items. Their location among such items is indicative of their temporal extent. We show a small subset of `VocalSound` types in Table 1. As can be seen, the `VocalSound` type `laugh` is the most frequently annotated non-verbal vocal production. The second type of laughter-relevant annotation found in the corpus, `Comment`, describes events of extended duration which were not localized between specific lexical items. In particular, this annotation covers the phenomenon of "laughed speech" [13] We list the top five most frequently occurring `Comment` descriptions pertaining to laughter in Table 2. As with `VocalSound` descriptions, there is a large number of very rich laughter annotations each of which occurs only once or twice.

Table 1: Top 5 most frequently occurring `VocalSound` types in the ICSI Meeting Corpus, and the next 5 most frequently occurring types relevant to laughter.

Freq Rank	Token Count	`VocalSound` Description	Used Here
1	11515	`laugh`	√
2	7091	`breath`	
3	4589	`inbreath`	
4	2223	`mouth`	
5	970	`breath-laugh`	√
11	97	`laugh-breath`	√
46	6	`cough-laugh`	√
63	3	`laugh, "hmmph"`	√
69	3	`breath while smiling`	
75	2	`very long laugh`	√

The description attributes of both the `VocalSound` and `Comment` tags, as produced by the ICSI transcribers, appear to be largely ad hoc, and reflect practical considerations during an annotation pass whose primary aim is to produce an orthographic transcription. In the current work, we used the descriptions only to select and possibly segment laughter, and afterward ignored them.

We identified 12635 transcribed `VocalSound` laughter instances, of which 65 were ascribed to farfield channels. These were excluded from our subsequent analysis, because the ICSI MRDA Corpus includes forced alignment information for nearfield channels only. We also identified 1108 transcribed `Comment` laughter instances, for a total of 13678 transcribed laughter instances in the original ICSI transcriptions.

2.3. Laugh Bout Segmentation

Our strategy for producing accurate endpoints for the laughter instances identified in Subsection 2.2. consisted of a mix of automatic and manual methods. Of the 12570 non-farfield `VocalSound` instances, 11845 were adjacent on both the left and the right to either a time-stamped utterance boundary, or a lexical item. We were thus able to automatically deduce start and end times for 87% of the laughter instances treated in this work. Each automatically segmented instance was inspected by at least one of our annotators; disagreement as to the presence of laughter was investigated by both authors together, and in a small handful of cases ($<$3%), when there appeared to be ample counter-evidence,

Table 2: Top 5 most frequently occurring `Comment` descriptions containing the substring "`laugh`" or "`smil`". We listened to all utterances whose transcription contained these descriptions, but portions were included in our final laugh bout segmentation only if the utterances contained laughter (in particular, intervals annotated with "`while smiling`" were not automatically included.)

Freq Rank	Token Count	`Comment` Description
2	980	`while laughing`
16	59	`while smiling`
44	13	`last two words while laughing`
125	4	`last word while laughing`
145	3	`vocal gesture, a mock laugh`

we discarded the instance.

The remaining 725 non-farfield `VocalSound` instances were not adjacent to an available timestamp on either or both of the left and the right. These instances were segmented manually, by listening to the entire utterance containing them[1]; since the absence of a timestamp was due mostly to a transcribed, non-lexical item before and/or after the laughter instance, segmentation consisted of determining a boundary between laughter and, for example, throat-clearing. We did not attempt to segment one bout of laughter from another.

All of the 1108 `Comment` instances were segmented manually. This task was more demanding than manual segmentation of `VocalSound` laughter. We were guided by the content of the `Comment` description, which sometimes provided cues as to the location and extent of the laugh (ie. `last two words while laughing`). We placed laughter start points where the speaker's respiratory function was perceived to deviate from that during speech; in determining the end of laughter, we included the audible final recovery inhalation which often accompanies laughter [6].

A quarter of the manually segmented `Comment` instances were checked by the second author. The final laugh bout segmentation was formed by combining the automatically segmented `VocalSound` laughter, the manually segmented `VocalSound` laughter, and the manually segmented `Comment` laughter; due to overlap, a small number of laugh segments were merged, to yield 13259 distinct segments.

We note that the resulting laugh bout segmentation differs from that recently produced for the same corpus in [20] at least in the number of bouts. The authors of [20] report using only 3574 laughter segments; it is unclear how these were selected, except that the authors state that they excluded speech and inaudible laughter after listening to all the ICSI-transcribed instances.

2.4. Laugh Bout Voicing Classification

In the last preprocessing task, we classified each laughter instance as either voiced or unvoiced. Our distinction of voiced versus unvoiced was made according to [14]. Voiced laughter, like voiced speech, occurs when the energy source is quasi-periodic vocal-fold vibration. This class includes melodic, "song-like" bouts, as well as most chuckles and giggles. Unvoiced laughter results from fricative excitation, and is analogous to whispered speech. It includes open-mouth, pant-like sounds, as well as closed-mouth grunts and nasal snorts. Additionally, we decided that bouts consisting of both voiced and unvoiced calls should receive the voiced label when taken together. Instances of "laughed speech" were automatically assigned the voiced label.

Voicing classification was performed by two annotators, who were shown all the close-talk channels per meeting in parallel, for all segmented instances of laughter from Subsection 2.3. with their original ICSI `VocalSound` or `Comment` annotation. For each instance, they were able to select and listen to the foreground channel, the same time interval on any of the remaining channels, and the temporal context on the foreground and remaining channels[2]. Annotators were encouraged to insert ad-hoc comments in addition to their voiced/unvoiced label.

58 meetings were labeled by one of two annotators, 14 were labeled by one annotator and were then checked by the other, and 3 were independently labeled by both annotators. Finally, all laughter instances which received a comment during classification were subsequently listened to by both authors.

Interlabeler agreement on the classification of voicing was computed using the three meetings which were labeled independently by both annotators, `Bmr016`, `Bmr018` and `Bmr019`. Agreement was between 88% and 91%, and chance-corrected κ-values [5] for the three meetings fell in the range 0.76-0.79. This is lower than we expected, having had assumed that assessment of voicing is not a very subjective task. Inspection of the disagreements revealed that they occurred for `VocalSound` instances whose endpoints had been inferred from inaccurate forced alignment timestamps of the adjacent words. In many cases the annotators had labeled the presence of laugher speech inside laugh bouts; since commented cases were revisited by both authors, a portion of the disagreement cases were resolved. In the remainder, we kept the voicing label

of that of our two annotators who had worked on the larger number of meetings.

In a final verification effort (following the publication of [11]), the second author checked the voicing label and boundaries of every instance, which led to a change of voicing label in 942 instances. Endpoints were modified in 306 instances, and 50 instances were removed. 11961 laughter segments (90% of the total) were not modified.

3. ANALYSIS

In this section, we describe the results of our investigations into the differences between voiced and unvoiced bouts of laughter, in terms of total time spent in laughter, bout duration, and multiparticipant vocal activity context.

3.1. Quantity

Of the 13209 bouts identified in the previous section, 33.5% were labeled as unvoiced while 66.5% were labeled as voiced.

We were also interested in the total proportion of time spent laughing. For each participant, and for each of voiced and unvoiced laughter categories, we summed the time spent laughing, and normalized this quantity by the total time of those meetings which were attended by that participant. Since a given participant may not have been present for the entirety of each meeting, the results we show represent ceiling numbers.

We found that the average participant spends 0.98% of their total meeting time in voiced laughter, and 0.35% of their total meeting time in unvoiced laughter. For contrast, in [11], we showed that the average participant spends 14.8% of their total meeting time on speaking. It can be seen in Figure 1, that the time spent laughing and the proportion of voiced to unvoiced laughter vary considerably from participant to participant.

Visually, there appears to be only a very weak correlation between the amount of individual participants' voiced laughter and their amount of unvoiced laughter. The majority of participants appears capable of both modes of laughter production.

3.2. Duration

Next, we analyze the durations of bouts to determine whether there is a difference for voiced and unvoiced laughter. The results are shown in Figure 2. Although bout durations vary much less than talkspurt durations, the modes for all three of voiced laughter bouts, unvoiced laughter bouts, and talkspurts fall between approximately 1 second and 1.5 seconds. On average, voiced bouts appear to be slightly longer than unvoiced bouts.

Figure 1: Proportion of total recorded time per participant spent in voiced and in unvoiced laughter. Participants are shown in order of ascending proportion of voiced laughter.

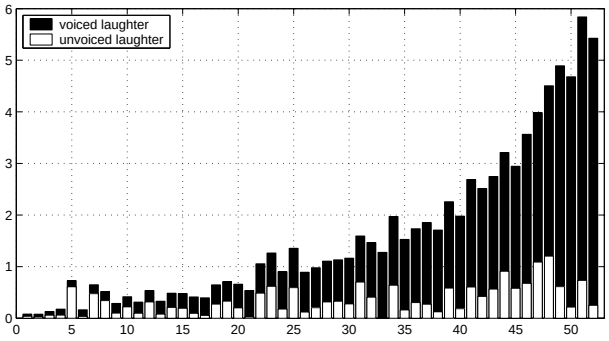
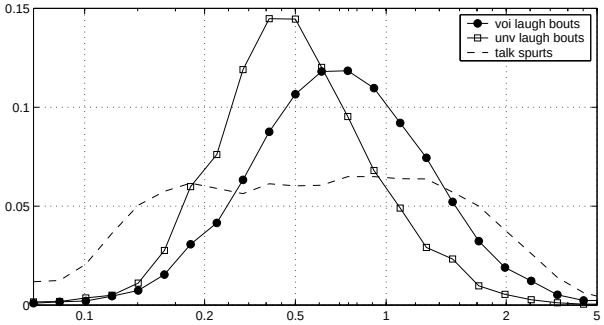


Figure 2: Normalized distributions of duration in seconds for voiced laughter bouts, unvoiced laughter bouts, and talk spurts.

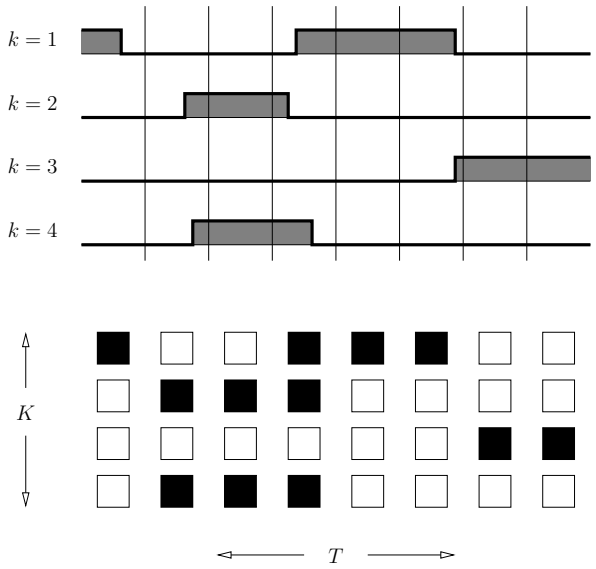


3.3. Interaction

Finally, we attempt to analyze local (short-time) differences in conversational context for voiced and unvoiced laughter. We are interested in whether a choice of voicing during laughter has a significant impact on the kinds of vocal interaction which immediately follow, or whether preceding interaction has a significant impact on whether a laugher will employ voicing. For each bout, we study only the *vocal interaction* context; in particular, we ignore the specific words spoken and focus only on whether each participant is silent, laughing (in either voiced or unvoiced mode), speaking, or both.

We accomplish this analysis in a time-synchronous fashion as follows, accumulating statistics over all meetings in the ICSI corpus. For every meeting, we begin with the reference on-off patterns corresponding to speech (Subsection 2.1.), for each of K participants. We discretize these patterns using 1-second non-overlapping windows, as shown in Figure 3. We do the same with the

Figure 3: Discretization of multichannel speech (or voiced or unvoiced laughter) segmentation references using a non-overlapping window size of 1 second. When participant k is vocalizing for more that 10% of the duration of frame t, the cell (t, k) is assigned the value 1 (black); otherwise it is assigned 0 (white).

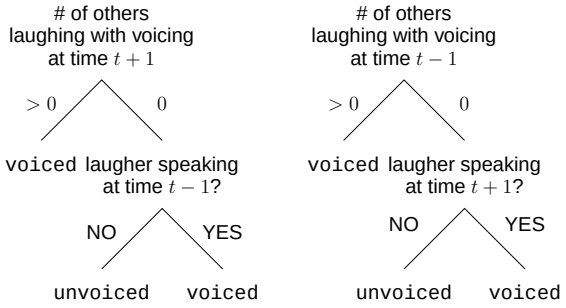


on-off voiced laughter segmentation and the on-off unvoiced laughter segmentation, producing for each meeting 3 binary-value matrices of size $K \times T$, where T is the number of 1-second non-overlapping frames.

For each meeting in the corpus, we inspect the reference matrices described above to determine whether participant k is laughing at time t. If so, we collect 11 features describing the conversational context of cell (t, k): binary-valued features whether participant k, the laugher, is speaking at times $t - 1$ and $t + 1$; the number of *other* participants speaking at times $t - 1$, t, and $t + 1$; the number of *other* participants laughing with voicing at times $t - 1$, t, and $t + 1$; and the number of *other* participants laughing without voicing at times $t - 1$, t, and $t + 1$.

We wish to analyze interactional aspects during laughter initiation, laughter termination, and laughter continuation, separately. To determine whether voiced and unvoiced bouts of laughter differ in terms of their short-time conversational context during initiation, we take all laughter frames which are preceded immediately by not-laughter, from all meetings and all participants, and measure the statistical significance of association between the 11 features we collect and the binary voiced or unvoiced attribute of the laughter frame. Although a standard χ^2-test is possible, we choose instead to determine whether *the voicing attribute during laughter initiation is predictable from context*. We do this by inferring the parameters of a decision tree [16], followed by pruning. Tree nodes which survive pruning are statistically significant; structurally, the decision tree can be thought of as a nested χ^2-test.

Figure 4: Automatically identified decision trees for detecting voiced versus unvoiced laughter based on multiparticipant vocal activity context; laughter initiation context on left, termination context on right.



We repeat the same procedure for both laughter termination and for laughter continuation. Our experiment identifies no significant distinction between the conversational context of voiced and unvoiced laughter continuation. That is, there appears to be no significant difference in the kinds of interactions that occur during voiced and unvoiced laughter, nor does voicing during laughter appear to have a significant impact on the interactions that occur during it.

For initiation and termination frames, we show the inferred classification trees in Figure 4. It is surprising that the two trees are symmetric. In attempting to predict the voicing of a frame which initiates a bout, the most useful contextual feature, of those studied here, is whether others will be laughing at $t + 1$; in other words, voicing during laughter is significantly more likely to cause at least one other participant to subsequently laugh with voicing. In attempting to predict the voicing of a frame which terminates a bout, the most useful feature is whether others were laughing with voicing at $t - 1$. Again, this suggests that voicing during laughter is much more likely if others were previously laughing with voicing. The next most useful feature is whether the laugher is speaking before or after laughing. For bout-initiating frames, if no others subsequently laugh with voicing and the laugher was not previously speaking, they are much more likely to be lau-

ghing without voicing, and symmetrically for bout-terminating frames.

4. CONCLUSIONS

We have produced a complete voiced and unvoiced laughter segmentation for the entire ICSI Meeting Corpus, including isolated instances as well as instances of laughter co-occurring with the laugher's speech. We have shown that on average, voiced laughter accounts for 66.5% of all observed laughter in this corpus, but that participants vary widely in their use of voicing while laughing. Most importantly, we have shown that in spite of inter-participant differences, voiced and unvoiced laughs are correlated with different vocal interaction contexts. Voiced laughter seems to differ from unvoiced laughter in that voiced laughter from other participants follows its initiation and precedes its termination. Voiced laughter also seems more interdependent with the laugher's speech; in cases where laughter follows speech or precedes laugher's speech, it is more likely to be voiced than unvoiced.

5. ACKNOWLEDGMENTS

This work was funded in part by the European Union under the integrated project CHIL (IST-506909), Computers in the Human Interaction Loop (`http://chil.server.de`) [21].

6. REFERENCES

[1] Bachorowski, J.-A., Smoski, M., Owren, M. 2001. The acoustic features of human laughter. *J. Acoustical Society of America* 110(3), 1581–1597.

[2] Burger, S., MacLaren, V., Yu, H. 2002. The ISL Meeting Corpus: The impact of meeting type on speech style. *Proc. ICSLP* Denver CO, USA. 301–304.

[3] Carletta, J. e. a. 2005. The AMI Meeting Corpus: A pre-announcement. *Proc. MLMI (Springer Lecture Notes in Computer Science 3869)* Edinburgh, UK. 28–39.

[4] Chafe, W. 2003. *Linguistics, Language, and the Real World: Discourse and Beyond* chapter Laughing while Talking, 36–49. Georgetown University Press.

[5] Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 37–46.

[6] Filippelli, M. e. a. 2001. Respiratory dynamics during laughter. *J. Applied Physiology* 90(4), 1441–1446.

[7] Janin, A. e. a. 2003. The ICSI Meeting Corpus. *Proc. ICASSP* Hong Kong, China. 364–367.

[8] Jefferson, G. 1979. *Everyday Language: Studies in Ethnomethodology* chapter A Technique for Inviting Laughter and its Subsequent Acceptance Declination, 79–96. Irvington Publishers.

[9] Jefferson, G. 1985. *Handbook of discourse analysis* volume 3 chapter An exercise in the transcription and analysis of laughter, 25–34. Academic Press.

[10] Laskowski, K., Burger, S. 2006. Annotation and analysis of emotionally relevant behavior in the ISL Meeting Corpus. *Proc. LREC* Genoa, Italy.

[11] Laskowski, K., Burger, S. 2007. Analysis of the occurrence of laughter in meetings. *Proc. INTERSPEECH (to appear)* Antwerpen, Belgium.

[12] Neiberg, D., Elenius, K., Laskowski, K. 2006. Emotion recognition in spontaneous speech using GMMs. *Proc. INTERSPEECH* Pittsburgh PA, USA. 809–812.

[13] Nwokah, E., Hsu, H.-C., Davies, P., Fogel, A. 1999. The integration of laughter and speech in vocal communication: a dynamic systems perspective. *J. Speech, Language & Hearing Research* 42, 880–894.

[14] Owren, M., Bachorowski, J.-A. 2003. Reconsidering the evolution of nonlinguistic communication: The case of laughter. *J. Nonverbal Behavior* 27(3), 183–199.

[15] Provine, R. 1992. Contagious laughter: Laughter is a sufficient stimulus for laughs and smiles. *Bull. Psychonomic Society* (30), 1–4.

[16] Quinlan, J. 1986. Induction of decision trees. *Machine Learning* 1(1), 81–1006.

[17] Shriberg, E. 2005. Spontaneous speech: How people really talk, and why engineers should care. *Proc. INTERSPEECH* Lisbon, Portugal. 1781–1784.

[18] Shriberg, E., Stolcke, A., Baron, D. 2001. Observations on overlap: Findings and implications for automatic processing of multi-party conversation. *Proc. EUROSPEECH* Aalborg, Denmark. 1359–1362.

[19] Shriberg, E. e. a. 2004. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. *Proc. SIGdial* Cambridge MA, USA. 97–100.

[20] Truong, K., van Leeuwen, D. February 2007. Automatic discrimination between laughter and speech. *Speech Communication* 49(2), 144–158.

[21] Waibel, A., Steusloff, H., Stiefelhagen, R. 2004. CHIL: Computers in the Human Interaction Loop. *Proc. ICASSP2004 Meeting Recognition Workshop* Montreal, Canada.

[1] We used the freely available Audacity© for this task. Only the foreground channel for each laughter instance was inspected.

[2] We used our in-house multichannel annotation tool TransEdit for this task

Whom we laugh with affects how we laugh

Nick Campbell

NiCT/ATR-SLC
National Institute of Information and Communications Technology
& ATR Spoken Language Communication Research Labs
Keihanna Science City, Kyoto 619-0288, Japan
nick@nict.go.jp, nick@atr.jp

ABSTRACT

This paper describes work that shows how the acoustic features of laughter in Japanese speech vary according to conversational partner, reflecting the social status of laughter, and confirming that even such a simple sound is affected by non-linguistic factors such as social or intercultural relationships. Neural networks were successfully trained to identify the nature of the interlocutor from principal components of the acoustic and prosodic features of the laughing speech.

Keywords: Laughter, laughing speech, voice quality, acoustic characteristics, principal-component analysis, neural-network training

1. INTRODUCTION

The acoustics of laughter have been shown to be both highly complex and highly variable [1] with voiced and unvoiced variants functioning separately and having different effects [2]. However, most studies of laughter have been concerned with reactions to media rather than with laughter in interactive conversational situations. Recent work by [3] has shown laughter in conversation to be much more frequent than has been described previously in the literature, and suggests that this form of interactive laughter may primarily serve both to regulate the flow of the interaction and to mitigate the meaning of a preceding utterance. High intra-individual variability which greatly exceeded the parameter variability between subjects was found in the acoustic parameters of this type of laughter. The present paper extends this work to examine how laughter in the speech of two Japanese adults also varies systematically according to the nature of the interlocutor.

In this paper we make use of a global measure of the acoustics of laughter, derived from a principal component analysis of fourteen basic measures of prosodic and spectral characteristics incorporating voice quality [4]. It has been shown elsewhere [5] that this measure correlates closely with the changes in speaking style that occur with differences in familiarity between a speaker and a listener, and with differences in the ease of conversation that arise from e.g., cross-cultural or cross-language interactions. In the present paper we examine the changes in these characteristics that occur in the laughter and laughing speech of two Japanese individuals, one man and one woman, in conversations with four strangers over a period of time. The speech is in Japanese, but it is likely that the phonetic and prosodic characteristics of laughter are common to all people of whatever language background. However, the nature and style of laughing may of course vary considerably according to cultural and situational constraints.

Table 1: Counts of utterances extracted from the corpus. All are laughs, those on the right are laughing while speaking. C and E represent Chinese and English native-language partners, F and M the sex of the interlocutor. The sex and language of the speaker is shown in the second row

	laughs		+ speech	
	JF	JM	JF	JM
CF	201	241	131	214
CM	174	174	93	156
EF	350	401	140	173
EM	228	232	100	122

2. DATA

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 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 socialise 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 ten sessions between each pair. Each conversation lasted for a period of thirty minutes.

In all, ten people took part as speakers in the corpus recordings, five male and five female. Six were Japanese, two Chinese, and two native speakers of American English. All conversations were held in Japanese. There were no constraints on the content of the conversations other than that they should occupy the full thirty-minute time slot. Partners were initially strangers to each other, but became friends over the period of the recordings. The conversations between the three pairs of Japanese speakers form the main part of this corpus [5], and the conversations with non-native speakers form a subpart which is reported here. The non-native speakers were living and working in Japan, competent in Japanese, but not at a level approaching native-speaker fluency.

The speech data were transferred to a computer and segmented into separate files, each containing a single utterance. Laughs were marked with a special diacritic, and laughing speech was also bracketed to show by use of the diacritic which sections were spoken with a laughing voice. Laughs were transcribed using the Japanese Katakana orthography, wherever possible, alongside the use of the symbol.

The present analysis focusses on these two types of laughter as produced by the two Japanese speakers who spoke to the highest number of partners (see Table 1 for counts), and examines the changes depending on relationship with the interlocutor as characterised by native-language and sex.

3. MODELLING THE LAUGHS

Laughter was very common in the speech of all the conversation participants. Their situation was unusual in that although they did not initially know each other, they were required to talk over a telephone line (with no face-to-face contact) for a period of thirty minutes each week for five weeks. They were all paid and willing volunteers and knew that their recordings would be used for telecommunications research, but they had no detailed knowledge about the purpose of the recordings. Over the period of five conversations, they came to know each other quite well.

The transcribed speech files containing laughter were processed by a computer program to extract a set of acoustic features for each utterance. Since the utterances were typically short, we used a single value for each feature to describe an utterance. The features included pitch, power, duration, and spectral shape. Pitch was described by the mean, maximum, minimum, location of the peak in the utterance, and degree of voicing throughout the utterance. Power was described by the mean, maximum, minimum, and location of the peak in the utterance. Duration of the whole utterance was expressed as a log value, and a simple estimate of speaking rate was made by dividing the duration by the number of moraic units in the transcription. Spectral shape was described by the location and energy of the first two harmonics, the amplitude of the third formant, and the difference in energy between the first harmonic and the third formant (h1-a3, proposed by Hansen as the best measure for describing breathiness in her study of the voice quality of female speakers [6]). All these measures were produced automatically using the Tcl/Tk "Snack" audio processing library [7]. Thus for each laughing utterance in the conversations, we produced a vector of values corresponding to its acoustic characteristics.

Table 2: Cumulative proportion of the variance accounted for by the principal component analysis. 'f' and 'm' stand for female and male, and 'l' and 's' for laughter and laughing-speech respectively. Only the first 10 components are shown

	pc1	pc2	pc3	pc4	pc5	pc6	pc7	pc8	pc9	pc10
f-l	.23	.43	.54	.64	.72	.78	.84	.88	.92	.95
m-l	.31	.45	.57	.66	.74	.79	.84	.89	.92	.95
f-s	.21	.35	.49	.58	.65	.72	.79	.85	.89	.93
m-s	.19	.33	.46	.55	.64	.72	.78	.84	.89	.93

3.1. Principal Component Analysis

To simplify the use of these acoustic features in training a statistical model, we performed a principal component analysis [8] using the "princomp" function call in R [9]. The first three principal components account for about 50% of the variance in the acoustic data, and the first seven components together account for more than 80%. Table 2 shows that the first five principal components accounted for approximately 73% of the acoustic and prosodic variance in the laughs, and approximately 65% of the acoustic and prosodic variance in the laughing speech. The limited phonetic component of simple laughter makes it acoustically less variable than the laughing speech, and hence slightly easier to model.

3.2. Neural Network Training

In order to determine whether the variance observed in these laughs was related in any way to the nature of the interlocutor, a neural network was trained to learn the mapping between the first five (5) principal components and a label representing either (a) Chinese vs. English, or (b) male vs. female.

Table 3: Raw scores for the neural network trained to distinguish between Chinese and English-speaking partners. Here, C stands for Chinese, and E for English, with X for indeterminate ('don't know') predictions.

laughs						
	JF			JM		
→	E	C	X	E	C	X
CF	34	304	162	40	338	122
CM	41	299	160	33	350	117
EF	326	47	127	318	56	126
EM	284	45	171	350	30	120
laughing speech						
	JF			JM		
→	E	C	X	E	C	X
CF	53	353	94	34	369	97
CM	39	383	78	49	351	100
EF	369	42	89	358	45	97
EM	388	38	74	363	37	100

Table 4: Raw scores for the neural network trained to distinguish between male and female partners. Here, M stands for male, and F for female, with X for indeterminate predictions. In all cases, the 'correct' answer predominates.

laughs						
	JF			JM		
→	M	F	X	M	F	X
CF	71	259	170	34	337	129
CM	271	24	205	318	59	123
EF	14	352	134	26	349	125
EM	277	22	201	326	53	121
laughing speech						
	JF			JM		
→	M	F	X	M	F	X
CF	39	333	128	37	341	122
CM	352	26	122	358	43	99
EF	43	324	133	46	329	125
EM	332	40	128	353	27	120

A back-propagation neural network was constructed with five input neurons representing the activation of the first five (5) principal components of the acoustics of the laughter, or laughing speech, with a layer of seven (7) intermediate neurons and an output layer of two (2) neurons representing either male or female partners, or Chinese native-language or English native-language partners depending on the training session. The $nnet$ function of R was used for this with the following arguments:
$pcnet = nnet.formula(who \sim pc1 + pc2 + pc3 + pc4 + pc5, size = 7, rang = 0.1, decay = 5e-4, maxit = 500, trace = F)$
and repeatedly trained for each combination of speaker, laughing type, and interlocutor pattern.

We randomly selected from the utterances shown in Table 1 a subset of fifty (50) tokens for each partner of male and female laughter and laughing speech samples for training (giving 4×200 tokens in all) and a separate set of 50 each for testing in each category. Using an arbitrary threshold, values greater than 0.5 in the output neurons were taken as positive, less than -0.5 as negative, and values between -0.5 and 0.5 were taken to indicate that the network could not distinguish between training classes on the basis of the five principal component values for each token.

The network was trained with fifty (50) samples each of (a) laughter and (b) laughing speech randomly selected from conversations with each class of partner (c,d), giving a training vector of two-hundred ($4 \times 50 = 200$) samples. The trained network was then tested on a completely different vector of two-hundred (200) samples from a different random selection under the same criteria.

Because the networks are randomly initialised, and can produce different results with each training session, we performed ten (10) training and testing cycles for each combination and summed the results for each prediction category. These are the figures reported in the Tables. Tables 3 and 4 give the raw training results for each combination. The labels 'E', 'C', and 'X' in Table 3 indicate predictions for English, Chinese and 'don't-know' for Chinese female partner (CF), Chinese male partner (CM) etc. It can be seen from the tables that the networks successfully identify the partner from the acoustics of the laughter or laughing speech in the majority of cases.

4. RESULTS

Tables 5 and 6 show expanded summaries of the data in Tables 3 and 4 for a comparison of differences between the various prediction tasks. Statistics for the networks trained to detect the sex of the partner from the rotated acoustic parameters are shown in Table 5, and those for the Chinese/English discrimination in Table 6. The two leftmost columns in the tables provide summed results, disregarding individual partner differences (which can be examined from Tables 3 or 4). No test is necessary to see that these differences are significant, with more than six hundred correct responses against less than a hundred false responses in every case.

The centre two columns of the table are more revealing. They show counts of hits, misses, and

Table 5: Summed scores for the trained networks predicting male/female partner distinction from acoustic parameters. Indeterminate prediction results are shown in brackets, see text for an explanation.

laughs - JF					
discrim.		accuracy		JF	
85	611	131	1159	279	2500
548	46	-	(710)	-	(1221)
laughing speech - JF					
discrim.		accuracy			
82	657	148	1341	-	-
684	66	-	(511)	-	-
laughs - JM					
discrim.		accuracy		JM	
60	686	172	1330	325	2711
644	112	-	(498)	-	(964)
laughing speech - JM					
discrim.		accuracy			
83	670	153	1381	-	-
711	70	-	(466)	-	-

Table 6: Summed scores for the trained networks predicting Chinese/English partner distinction from acoustic parameters. Indeterminate prediction results are shown in brackets, see text for an explanation.

laughs - JF					
discrim.		accuracy		JF	
75	603	167	1213	339	2706
610	92	-	(620)	-	(955)
laughing speech - JF					
discrim.		accuracy			
92	736	172	1493	-	-
757	80	-	(335)		
laughs - JM					
discrim.		accuracy		JM	
73	688	159	1356	324	2797
668	86	-	(485)	-	(879)
laughing speech - JM					
discrim.		accuracy			
83	720	165	1441	-	-
721	82	-	(394)	-	-

'don't-know' responses for each class of speaker and laughing style. The two columns on the right of the table summarise these figures across each style of laughter to provide overall scores for each speaker.

Pearson's Chi-Square test [10] was used to compare each pair of results, and only JF(m/f) and JM(m/f) showed any significant differences.
JF-M/F accuracy (85, 611, 46, 548):
$\rightarrow \chi^2$ = 6.53, df = 1, p = 0.01059 (signif)
JF-M/F confidence (611, 304, 548, 406):
$\rightarrow \chi^2$ = 16.87, df = 1, p = 3.987e-05 (signif)
JM-M/F accuracy (60, 686, 112, 644):
$\rightarrow \chi^2$ = 16.32, df = 1, p = 5.349e-05 (signif)
JM-M/F confidence (686, 254, 644, 244):
$\rightarrow \chi^2$ = 0.02, df = 1, p = 0.8678 (n.s.).

No other differences between correct and false partner discriminations are significant. However, the difference in performance for JF overall, comparing discrimination success, is considerable:
JF-M/F overall (964, 2711, 1221, 2500):
$\rightarrow \chi^2$ = 38.17, df = 1, p = 6.478e-10 (signif)
cf JM-M/F overall (955, 2706, 879, 2797):
$\rightarrow \chi^2$ = 4.51, df = 1, p = 0.03373 (n.s.).

However, even for the least successful case (predicting the sex of the interlocutor from the style of JF's laughter) the network achieves 62.5% accuracy against a chance score of 50%. The male speaker's laughing idiosyncrasies allow the network to predict the sex of his interlocutor at 67.7% accuracy. The female speaker, differentiates her style of laughter when talking with foreigners sufficiently for the network to discriminate at 67.5%, and the male at an even higher rate of 70%.

From these stringent training and testing conditions one can conclude that the network is indeed able to generalise from the features of the acoustics in order to be able to identify the interlocutor at rates significantly better than chance. This confirms that speakers modify their laughter in a consistent way that indicates something about the nature of their relationship with the interlocutor.

5. DISCUSSION

JM laughs most with CF and EF; JF laughs least with CM and EM. Both laugh much more with EF (whose Japanese is less than fluent). It remains as future work to examine the nature of those relationships and the role of the individual acoustic features in triggering the different perceptions. However, some details of the acoustic mapping are given in Table 7 which shows first three principal components in each situation. The numbers are related to the strength of contribution of each acoustic feature in each component. For simplicity, values lower than 25 have been replaced by dashes to facilitate comparison. The table shows that in all cases the breathiness of the voice, as indicated by h1-a3 (a measure of spectral tilt, derived from subtracting energy measured at the third formant from energy measured at the first harmonic) plays an important contribution with strong weightings in every case.

Table 7: Contribution values (rotations) of each prosodic or acoustic feature in the first three principal components of each speaker-laughing-style combination. Vertical bars separate the laughing styles, and values for pc1, pc2, pc3 are listed in order within each. Values less than 25 have been replaced by dashes for simplicity

	JF-laugh			JF l+sp			JM-laugh			JM-l+sp		
fmean	--	35	--	45	--	--	33	--	--	32	32	--
fmax	29	37	--	--	27	--	30	27	--	36	--	34
fmin	39	--	--	34	--	28	33	--	--	--	--	--
fpct	--	--	--	--	--	28	--	--	39	--	--	32
fvcd	26	41	--	--	35	--	26	37	--	33	26	--
pmean	--	46	--	--	51	35	30	39	--	39	46	--
pmax	37	--	--	--	46	--	38	--	--	36	33	--
pmin	--	36	25	--	--	42	--	49	--	--	30	--
ppct	19	--	36	28	--	33	--	30	--	--	--	--
h1h2	--	--	25	--	--	--	--	--	--	--	--	35
h1a3	40	--	44	41	27	28	32	--	48	32	40	38
h1	40	--	26	38	--	36	33	--	32	32	39	--
a3	--	--	45	--	--	--	--	--	49	--	--	45
dn	--	31	37	34	--	30	--	47	--	--	--	29

The speakers control their voices differently, both in simple laughter and in laughing speech, and the differences in pitch, loudness and tension of the voice, or breathiness, reveal characteristics related both to the sex of the interlocutor and to differences in cultural background.

6. CONCLUSION

This paper has described a brief study of laughs and laughing speech excised from the telephone conversations of two Japanese speakers talking with two male and two female partners. It presented results showing that the speakers vary their laughing styles according to the sex and nationality of the partner.

A neural network was trained to distinguish either the sex of the interlocutor or their social background, as characterised by native language, and differences in the success of the training were compared for each of these two dimensions and for each of the two speakers.

It was shown in previous work [11] that a speaker adapts her voice quality as well as speaking styles according to the nature of her relationship with the interlocutor. The present study provides additional evidence for this common-sense but largely unexplored phenomenon by showing that differences can be also be found in the types of laughter expressed by a further two male and female speakers of Japanese in telephone conversations with four partners each over a period of five weeks.

In separate work with a very large single-speaker corpus [12] we found that approximately one in ten utterances contains laughter. From among these laughing utterances, we were able to distinguish four types of laughter according to what each revealed about the speaker's affective state, and were able to recognise these different types automatically by use of Hidden Markov Models trained on laugh segments, with a success rate of 75%. In future work we will attempt a similar perceptual classification of the different types of laughter found in the present corpus, and will attempt to explain their interpretation in a social and discourse context.

Acknowledgement

This work is partly supported by the National Institute of Information and Communications Technology (NiCT), and includes contributions from the Japan Science & Technology Corporation (JST).

7. REFERENCES

[1] Bachorowski, J.-A., Smoski, M.J., & Owren, M.J. (2001). "The acoustic features of human laughter", Journal of the Acoustical Society of America, 110, pp.1581-1597.

[2] Bachorowski, J.-A., and Owren, M.J. (2001). "Not all laughs are alike; Voiced but not unvoiced laughter elicits positive affect in listeners", Psychological Scienece 12, pp.252-257.

[3] Vettin, J. & Todt, D. (2004): "Laughter in conversation: features of occurrence and acoustic structure". J. Nonverbal Behaviour. 28: 93-115.

[4] Ni Chasaide, A., and Gobl, C., (1997). "Voice source variation". In W. J. Hardcastle and J. Laver (Eds.), *The Handbook of Phonetic Sciences*, pp. 428-461. Oxford: Blackwells.

[5] Campbell, N., (2007) "Changes In Voice Quality with respect to Social Conditions", in Proc 16th ICPhS 2007.

[6] Hanson, H. M., (1995). "Glottal characteristics of female speakers". Ph.D. dissertation, Harvard University.

[7] Käre Sjölander, (2006). The Snack Sound Toolkit from http://www.speech.kth.se/snack/

[8] Pearson, K., (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space". Philosophical Magazine 2 (6): 559-572.

[9] R Development Core Team, (2004). "R: A language and environment for statistical computing". R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, http://www.R-project.org

[10] Chernoff H, Lehmann E.L.(1954). "The use of maximum likelihood estimates in χ^2 tests for goodness-of-fit". Annals of Mathematical Statistics, 25:579-586.

[11] Campbell, N., and Mokhtari, P., (2003). "Voice Quality is the 4th Prosodic Parameter". *Proc. 15th ICPhS* Barcelona, pp.203–206.

[12] Campbell, N., Kashioka, H., Ohara, R., (2005). "No laughing matter", pp.465-468 in Proc INTERSPEECH-2005.