No Laughing Matter

Nick Campbell †*, *Hideki Kashioka* ‡*, *Ryo Ohara* *

Nara Institute of Science & Technology*
ATR Media Information Science Laboratories†
ATR Spoken Language Translation Laboratories‡
Keihanna Science City, Kansai, Japan

nick@atr.jp, hideki.kashioka@atr.jp, ryo-o@is.naist.jp

Abstract

Laughter matters! From an analysis of a very large corpus of naturally-occurring conversational speech we have confirmed 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 we were able to recognise these different types automatically, by use of Hidden Markov Models trained on laugh segments, with a success rate of greater than 75%. The paper also presents a speech synthesis interface that enables the control of such emotional expression for use in an interactive conversation.

1. Introduction

In order to discover what the more likely distributions of affective or emotional expressions might be, we produced a corpus of everyday conversational speech, which has been reported in detail elsewhere [1, 2]. In order to overcome Labov's well-known Observer's Paradox, wherein the presence of an observer or a recording device influences the productions of the observed person, our subjects agreed to wear small head-mounted studio-quality microphones for extended periods while going about their normal everyday social interactions over a period of about five years.

These volunteers were paid by the hour of speech that they produced, and a further group were paid to transcribe and annotate this speech data in fine detail. The transcriptions were produced in plain text rather than phonetic coding, but care was taken to transcibe every utterance exactly as it was spoken, with no effort made to 'clean-up' the transcriptions or correct the grammar.

Transcribers were encouraged to break the speech into the smallest possible utterance chunks by use of a notional yen-per-line payment policy. In spite of this, many single 'utterances' included several tens of syllables, being expressed as a single breath-group. The text of the transcriptions from one speaker, if printed end-to-end as a solid block of text in book form would fill 35 volumes, and if printed one-line-per-utterance, would probably exceed 100 volumes.

The majority of speech utterances in this corpus were single phrases; 'grunts', or phatic sounds made to reassure the listener of the speaker's affective states and discoursal intentions [3, 4]. Laughs were very frequent, as were back-channel utterances and fillers[1], but approximately half the number of utterances transcribed were unique. These typically longer utterances can perhaps be well handled by current speech synthesis techniques, since the text carries the brunt of the communication, but the shorter 'grunts' require a new method of treatment.

The word 'grunt' carries implications of pre-human or even animal behaviour, but perhaps this is the most appropriate term for the type of phatic communication that takes the place of mutual grooming in human society [5]. As well as the frequent "ummm", "ahhh", "yeah", "uh-uh", etc., We also include the use of such phrases as "good morning!" and "did you sleep well?", "see the game last night?", etc., which are used when social rather than propositional interactions are preferred. They float to the top of a multigram analysis [6] by dint of their frequent occurrence, but most can be characterised by the flexibility and variety of their prosody. None can be interpreted from the plain text alone. Perhaps these sounds are among the oldest forms of spoken language? In numerical terms, they account for more than half of the conversational corpus.

On the basis of the above distinction, we can now categorise the corpus utterances in terms of I-type and A-type functions; the former for the conveyance of information, the latter for the expression of affect [7, 8]. A framework was proposed (see Figure 1 for an illustration) which describes the two-way giving and getting of I-type and A-type information subject to speaker-state and listener-relationships. For simplicity in a speech synthesis application, we propose four levels of each:

- Self (the speaker herself)
  - Mood: the speech is 'brighter' if the speaker is in a 'good mood' (two levels: plus, minus).
  - Interest: the speech is more 'energised' if the speaker is interested in the conversation (two levels: high, low).
- Other (her relationships with the interlocutor)
  - Friend: the speech is 'softer' if the listener is a friend (two levels: close, distant).
  - Place: the speech is more 'intimate' if it takes place in a relaxed environment (two levels: reaxed, formal).

Any given utterance is realised subject to the above constraints. The challenge to synthesisers for conversational speech is to enable the user to specify such constraints. For A-type utterances, the framework is even more important than the text.

[1] We use the word 'filler', since it is common parlance, though we strongly object to the implication that a gap exists in the interaction which is being 'filled'. However, these slots in the communication process serve a very important function as places where non-linguistic (affective) communication can occur.

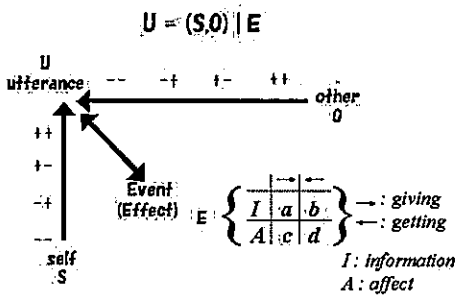


Figure 1: A framework for specifying the characteristics of an utterance according to speaker-state, relationship with the listener, and speech-act type.

2. Types of laughter

Many of the A-type utterances contain laughter. Looking specifically at the laughter in the ESP corpus, we find that 38,306 utterances were transcribed as laughing, while 398,655 were not. 11,331 were not further transcribed by the labellers, but the remainder were given a phonetic (using the Japanese kana alphabet) transcription. Table 1 lists a few of the more than 2000 types of laugh that occurred in the corpus. Table 2 extends this example by illustrating those transcribed laughs that had an (arbitrarily selected) occurrence of 5 tokens each in the corpus. Table 3 illustrates the most frequent 'words' that were uttered as a laugh.

In an attempt to categorise these laughs in a more systematic manner, we performed perceptual analyses of their characteristics and determined a basic set of fundamental laugh types. Four classes of laugh could be distinguished: a hearty laugh, an amused laugh, a satirical laugh, and a social laugh [9].

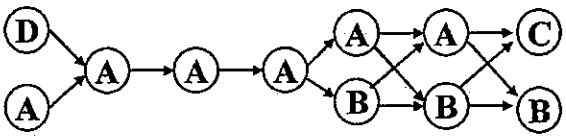
These functional laughter classes can be distinguished by differences in their segmental composition as described below. Here, we define a laugh 'segment' as a single sound-burst, a sequence of which usually makes up a single coherent laugh. Laughs can of course consist of only one segment, but usually they are both compound and complex.

Figure 2 illustrates a typical laughing sequence, with the upper part showing three of the four basic laughter segments that we have distinguished: A=voiced laugh, B=chuckle C=ingressive breathy laugh. The lower part of the figure completes the set with an example of a D=nasal grunt. These and further examples can be heard at http:feast.his.atr.jp/laughs.

A training set of 3000 laughs was hand-labelled both at the segment level and at the functional level by a specialist (the third author). Hidden Markov models were trained on multiple examples of each segment type to recognise them and output a sequence of types to characterise each laughing utterance. It was immediately noticed that they appear in regular sequences, with a grammar as illustrated below:

Table 1: Transcriptions and occurrence counts of the most frequent laughs in the ESP corpus. Note that some words (e.g., un, honma=really, aa) were also considered as laughs (see Table 2)

11331	untranscribed	98	[aha!]
409	[fufu]	97	[ha]
353	[ahaha]	92	[hahhahhahha]
307	[ahha]	86	[fuffufu]
299	[aha]	82	[ahahahahaha]
268	[ahhaha]	81	[hahahahaha]
266	[fufufu]	80	[fun]
261	[haha]	75	[ufufufu]
253	[hahaha]	75	[nfufu]
241	[ahahaha]	75	[hahhaha]
216	[fuffu]	74	[uffu]
213	[fufun]	74	[fufufufufu]
211	[haa]	72	[hahahahahaha]
197	[hahha]	68	[fuu]
189	[fu]	66	[ufun]
169	[ufu]	60	[ahahha]
167	[hahahaha]	59	[ffufu]
166	[fufufufu]	58	[fufu!]
166	<un>	58	[ahhahhahhahha]
147	[fuffuffu]	52	[ahhahahaha]
142	[ufufu]	61	<honma>
136	[ahahahaha]	48	<aa>
133	[hahhahha]	47	[hahhahhahhahha]
129	[ahhahha]	47	[haha!]
127	[ahhahhahha]	46	[uffufu]
107	[ahhahaha]	46	[hehe]
103	[fuffuffuffu]	46	[ehe]



Compared with the human-generated labels for each of the above basic-type sequences, the HMM recognised the segment sequences correctly at a rate of 81%. Further training using these derived segment labels to discriminate between the four functional laughter types listed above, through the use of the grammar, resulted in an overall success rate of 75% for the recognition of the functional types of each laugh (see [9] for full details).

Based on the the above results, we then used the derived HMMs to automatically label all the remaining laughs in the corpus.

3. Synthesising Laughter

Figure 3 shows a research prototype for an intention-based, clickable, conversational speech synthesis interface. 'Chakai'[2] allows for free input (by typing text into the white box shown at bottom-centre) as well as fast selection of various frequently-used phrases and, in addition, an icon-based speech-act selection facility for the most common types of 'grunt'. This format

[2]The name, not unrelated to CHATR is composed of two Japanese syllables, meaning tea-meeting, an event during which social and undirected chat is common.

Table 2: Laugh types with frequency 5 (arbitrarily chosen to further illustrate the phonetic variety of these utterances)

[ufu,ufu] [ufufu,fu] [ufu,fufu] [ufu,fu]
[uffufun] [uffufufufufu] [uffufuffu] [uffuffuffu]
[ohoho] [oho] [ohhoho] [ohho] [ntsu-] [nhahahaha]
[nfufufufufu] [ihi.] [hi.i.] [hehee] [he-] [hahhahahha]
[hahhahahahahaha] [hahha-] [hahahatsu]
[haha,hahahaha] [haha,hahaha] [hahahahaa]
[hahahaa] [haha-] [hafufufu] [haahaha] [ha,ahaha]
[funtsu] [funfunfun] [funfuffu] [fuhahahahaha]
[fufuu] [fufuhahahahahaha] [fufuhahaha] [fufuha]
[fufufufuffu] [fufufu,fuffu] [fu,fuffuffuffu]
[fuffufuffu] [fuffuffuffuffuffuffuffu] [ffuffu]
[ehheheheheh] [ehahahaha] [efu] [atsu,hahhahha]
[ahhahahhahha] [ahhaha,fuffu] [ahaha] [ahahahahha]
[ahahaha,fuffu] [aha,haha] [afufufufufu] [affutsu]

Table 3: Occurrence counts and transcriptions of the most frequent laughing words ('grunts') in the ESP corpus. Angle-brackets indicate the laughing portion

166	<un>	61	<homma>	48	<aa>	31	<ufun>		
28	<fun>	27	<e>	26	<a^>	23	<a>	22	<a,sounan>
18	<uso>	18	u<n>	17	<iya>	16	<nani>		
15	<sounan>	15	hon<ma>	15	<honde>				
15	<a,homma>	14	<nanka>	14	<fufun>				
14	<a,sou>	14	<ano>	13	<sousousousou>				

enables linking to a conventional CHATR-type synthesiser for creation of I-type utterances not found in the corpus, while providing a fast, three-click, interface for common A-type utterances which occur most frequently in ordinary conversational speech.

The selection of entire phrases from a large conversation-speech corpus requires specification not just of the intention of the phrase (a greeting, agreement, interest, question etc.,) but also of the speaker's affective state (as desired to be represented) and the speaker's long- and short-term relationships with the listener at that particular time. The icons that appear when a combination is selected indicate the available speech units in the corpus, and laughing variants are common. The corpus has been pre-annotated so the actual code that produces the segments is very simple. And since it is often the case that whole-phrase segments are concatenated with short pauses between them, the naturalness of the resulting speech can be absolute. No processing is required, thanks to the number and variety of utterances in the corpus.

Chakai can be used in real-time for conversational interaction. When initiating a topic, typed input is required, and this is presently too slow, but when showing interest or 'actively listening', then different grunts can be produced to encourage the speaker, challenge her, show surprise, interest, boredom, etc., by simply clicking on the icons.

The initial frame presents the user with a choice of four listener types: friend, family, stranger, or child, with adjustable bars for setting the activation of the Self and Other constraints. The following screen allows selection of different forms of greetings, sub-categorised according to occasion (e.g., morning, evening, telephone, face-to-face, initiation, reply etc.,) with an adjustable bar for setting the intended degree of activation or 'warmth of greeting' before the penultimate button-press. When these criteria are selected, the different types of speaking style representing available utterances in the corpus are presented as a row of activated smiley-faces (top of the figure) from which the user can select the closest to their intended interactional function.

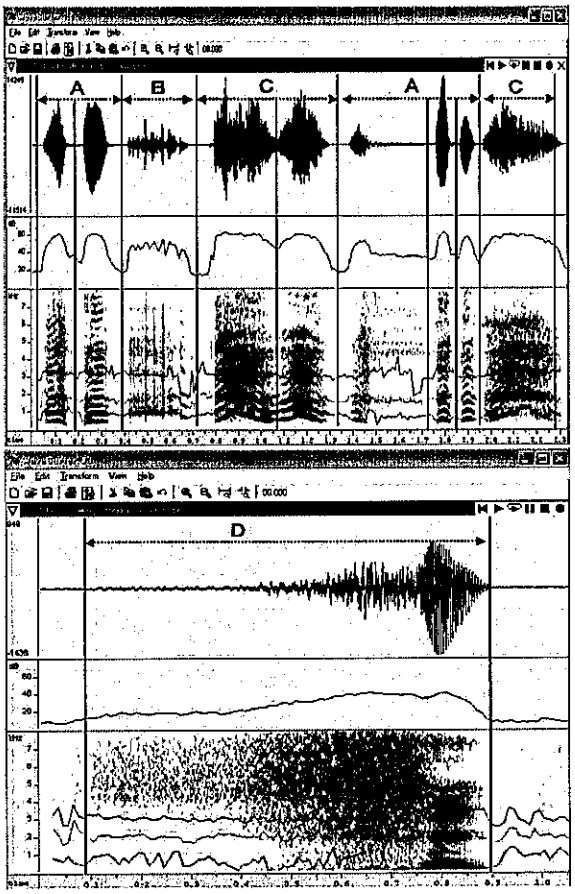


Figure 2: The syntax of laughter, top section: three types of laugh are concatenated in this natural-speech utterance. Lower section: a nasal grunt, type D

The following screen (shown in the figure) is for the core part of the conversational interaction. Icons are arranged in 4 rows, with questions on the right (who, where, why, when, etc.,) and positive, neutral, and negative grunts arranged in three columns on the left of the screen. The vertical dimension here is used for degree of activation.

By splitting utterances into three types, we have greatly facilitated the selection process. I-type utterances, largely unique since they are so content-dependent, have to be laboriously typed in; frequent phrases which are text-specific can be selected and a choice of speaking styles is offered via the smiley-face icon layer. Degree and type of of smile in the icon reflects the functional types of laughter in such utterances.

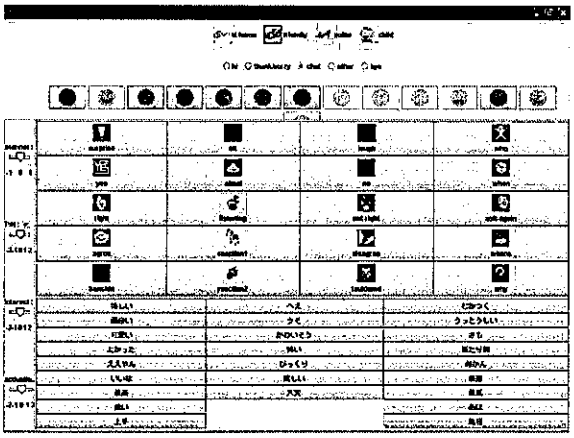


Figure 3: The Chakai Conversational Speech Synthesis interface. By clicking on a speech-act icon, a choice of emoticons is dispayed in the upper section of the screen according to corpus availability, from which an utterance having the appropriate speech characterstics can be selected. Utterances are selected at random from among those in that same category within the corpus so that subsequent selection of the same combination will provide natural variety without unnecessary repetition.

4. Summary and Conclusion

In this paper, we have introduced some of our recent work on the synthesis of conversational speech, and have shown that the challenges presented by this type of task are qualitatively different from those of conventional speech synthesis which is used primarily for the transmission of propositional content. We have found from our analysis of a very large natural-speech corpus that at least half of the utterances in interactive conversational speech are not well represented by their text alone and that they depend upon specific prosodic characteristics.

The paper has also described our initial attempts to classify laughter, a common form of interactive grunt, and to labeol it automatically for use in a speech synthesis device. We distinguished four types of function, and four types of phonetic segment to distinguish laughs, and showed that similar discrimination can be made by automatic methods for the efficient labelling of laught in conversational speech.

We described a prototype user-interface that allows input according to speech-act intention, using constraints representing the primary contextual influences on speaking-style, so that a conversational utterance can be produced rapidly with minimal input from the user. For the phatic utterances that are a characterisatic of informal and social speech, this interface allows text-free input, since an appropriate phrase is selected from the corpus according to the higher-level constraints automatically. This work is still experimental, and the paper should not be taken to imply that the methods presented here are necessarily the best for a commercial speech synthesis system, but it presents them as an illustration of the problem rather than of its solution. We are lucky to have such a speech corpus at our disposal, but replicating it for another language or subculture would require considerable extra work.

Clearly, this propotype does not represent the full final version, and it will require several generations of trial and evolution before an ideal conversation-device is realised, but we are satisfied that it well represents the problem that we are trying to solve. The user, whether handicapped or healthy, human or robot, should not have to specify the text of a conversational grunt, whether it be "yes" or "good morning" and then also have to describe its prosody or purpose. These are secondary characteristics of speech. They depend on the higher-level constraints of discourse context and speaker-intention just as the fine acoustic characteristics of CHATR segments depend on the phonetic and prosodic environment in which they occur. By knowing these dependencies and their interactions, we are able to simplify the process of selection and thereby to improve both the functionality and the quality of the synthesis process.

5. Acknowledgements

The first author gratefully acknowledges support from the ATR Network Informatics Lab, the Japan Science & Technology Agency, the National Institute of Information and Communications Technology (NiCT), and the Ministry of Public Management, Home Affairs, Posts and Telecommunications, Japan. The laughter analysis was carried out an NAIST by the third author.

6. References

[1] Campbell, N., "Recording Techniques for capturing natural everyday speech" pp.2029-2032, in Proc Language Resources and Evaluation Conference (LREC-02), Las Palmas, Spain, 2002

[2] Campbell, N., "Speech & Expression; the Value of a Longitudinal Corpus", pp.183-186 in Proc Language Resources and Evaluation Conference (LREC-04), Lisbon, Portugal, 2004.

[3] Campbell, N., & Erickson, D., "What do people hear? A study of the perception of non-verbal affective information in conversational speech", pp. 9-28 in Journal of the Phonetic Society of Japan, V7,N4, 2004.

[4] Campbell, N., "Specifying Affect and Emotion for Expressive Speech Synthesis", In, A. Gelbukh (Ed.) *Computational Linguistics and Intelligent Text Processing*, Proc. CICLing-2004. Lecture Notes in Computer Science, Springer-Verlag, 2004.

[5] Campbell,, N., "Getting to the heart of the matter; Speech is more than just the Expression of Text or Language", Keynote speech in Proc Language Resources and Evaluation Conference (LREC-04), Lisbon, Portugal, 2004.

[6] S. Deligne and F. Bimbot, "Language Modeling by Variable Length Sequences: Theoretical Formulation and Evaluation of Multigrams", pp.169-172 in Proc ICASSP, 1995.

[7] Campbell, N., "Listening between the lines; a study of paralinguistic information carried by tone-of-voice" pp 13-16, in Proc International Symposium on Tonal Aspects of Languages, TAL2004, Beijing, China, 2004.

[8] Campbell, N., "Extra-Semantic Protocols; Input requirements for the synthesis of dialogue speech", pp.221-228 in Andre E., Dybkjaer, L., Minker, W., & Heisterkamp, P., (Eds) *Affective Dialogue Systems*, Springer Lecture Notes in Artificial Intelligence Series, 2004.

[9] Ohara, Ryo, "Analysis of a laughing voice and the method of laughter in dialogue speech", unpublished Masters Thesis, Nara Institute of Science & Technology, 2004.