Conversational Speech Synthesis — and the need for some laughter

Nick Campbell
ATR Media Information Science Laboratories,
Keihanna Science City, Kyoto, 619-0288, Japan
nick@atr.jp

November 9, 2005

Abstract

This paper progress the synthesis of conversational speech, from the viewpoint of work carried out on the analysis of a very large corpus of expressive speech in normal everyday situations. With recent developments in concatenative techniques, speech synthesis has overcome the barrier of realistically portraying extra-linguistic information by using the actual voice of a recognisable person as a source for units, combined with minimal use of signal processing. However, the technology still faces the problem of expressing paralinguistic information, i.e., the variety in the types of speech and laughter that a person might use in everyday social interactions. Paralinguistic modication of an utterance portrays the speaker's affective states and shows his or her relationships with the speaker through variations in the manner of speaking, by means of prosody and voice quality. These inflections are carried on the propositional content of an utterance, and can perhaps be modelled by rule, but they are also expresssed through non-verbal utterances, the complexity of which may be beyond the capabilities of many current synthesis methods. We suggest that this problem may be solved by the use of phrase-sized utterance units taken intact from a large corpus.

Keywords Expression, Affect, Emotion, Social Interaction, Non-Verbal, Speech Synthesis, Conversation, Laughter

1 Introduction

The computer synthesis of natural-sounding speech has been a goal of computer scientists and speech technologists for more than half a century [1, 2], yet neither linguists nor phoneticians have yet achieved a comprehensive definition of the full range and variation of speech as a means of human communication and social interaction.

Most research into human language has been based on the analysis of written texts, and even when spoken language has been considered, it has been treated

either as a 'system of sounds' or as a 'media-transformed' version of text, to be analysed in written form through the use of transcriptions. This is understandable, since the technology for recording and analysing oral interactions has until recently been both expensive and lacking in portability. As a result, 'speech' is not well understood from the standpoint of 'communication'.

Similarly, conversation analysis has a long history of research, but again, in the majority of cases, it is the (usually cleaned-up) texts of the conversations that have formed the basic material for study. The actual sounds of the speech and their prosody have been considered as of secondary importance to the content; i.e., *What you say* has been treated as more important than *How you say it*; but whereas this may well be the case for information announcements, it is rarely so for casual conversational interactions, where phatic communion is as important as propositional content, if not even more so.

More recently, we find many comprehensive resources of spoken material available to researchers, thanks largely to the efforts of the speech recognition community to provide training material for their statistical engines. In the early days of speech recognition research the emphasis was more phonetic — categorising the basic speech sounds by use of Hidden Markov Models, and using triphone-contexts to define elemental phones, to be interpreted in conjunction with the use of a language model, in order to convert sound sequences into words for recognition. Prosodic variation in speech was considered irrelevant and largely ignored, because the technology provided word candidates regardless of the speaker-specic or utterance-specic variations. The texts could be understood without resource to prosodic knowledge, whch was thought to function primarily as a support for syntactic and semantic information already encoded in the text. Directed-speech, rather than casual conversation, was the norm in such research so the social uses of speech prosody were not realised.

The emphasis in speech data collection was on maximising speaker numbers in order to produce speaker-independent models, rather than on modelling the variations in the speech of a particular individual across time. Effects of differences in the listener were not considered important, as 'production rather than 'interaction' was the focus of the data collection. Developments in recognition technology were in the direction of whole-word modelling and in improvements to the statistical language models, but the assumption of a strong dependence between component phones and consequent word sequences remained. Recognition performance was and still is evaluated in terms of the number of words correctly transcribed. The assumption that the words alone can represent the speech has been largely unchallenged, and the fact that the same utterance can carry different meanings according to its pronunciations has been largely ignored, on the assumption that meaning can be understood from linguistic context alone.

Similarly for speech synthesis research, based on the early assumptions of synthesisers functioning as reading machines, the primary focus has been on the conversion of text sequences into sound sequences. From word-based input to speech output, the flow of processing is concentrated on predicting the sounds required to represent the word sequence in order to present the same

propositional content in a different medium. A given word is given different pronunciations depending on its context in an utterance, or on the syntactic structure of that utterance, but very little attention has yet been paid to the expression of affect or to the function of non-verbal utterances in speech.

Analysis of a very large corpus of natural conversational speech has shown that more than half of the utterances used in daily interaction have minimal propositional content and that they function instead to establish speaker-listener relationships and to express the speaker's affective states for phatic communication in way that cannot be transcribed into written text. This paper tackles the issue of how to synthesise such non-verbal, phatic utterances for use in conversational speech.

2 Corpus-based Speech Synthesis

Looking back across the long history of speech synthesis research, we can see in retrospect a clear evolution from the modelling of phonetic states to the modelling of utterance characteristics. The pioneering work of Gunnar Fant in Sweden [3, 4] and Dennis Klatt and his colleagues in the US [5, 6] was founded on a phonetic view of speech as a sequence of phones, modulated by prosody to represent syntactic and semantic content. Joe Olive [7, 8, 9], Osamu Fujimura, and their colleagues at Bell Labs made a signicant contribution by showing that the dynamics of the transitions between the phones carried much more information than an interpolated sequence of steady-state representations of phone centres. Yoshinori Sagisaka in Japan [10] extended this paradigm shift by concatenating non-uniform sequences of actual speech taken from readings of the most common 5000 words of the language. It became clear that the information carried in the dynamics of the speech far outweighed that of the supposed phonetic centres or steady states. The variation itself is the information that encodes the speech, and the art of synthesis lies in selecting the most appropriate variant, whether to reproduce it by rule or to reuse it in unit-selection.

Although text can be well represented by a sequence of invariant letters, speech sounds are not invariant. They depend heavily on the various contexts of their phonation [11]. My own work extended the above trend, showing that by incorporating prosodic contexts among the selection criteria for units for concatenation from a speech corpus [12, 13, 14], considerable improvement in speech realism (or information content) could be obtained. Although a small step in terms of unit-selection, this allowed us to remove the signal-processing component from the synthesiser and to use the speech segments intact, without resorting to potentially damaging signal modication. By simply concatenating phone-sized segments which had been selected according to both phonetic and prosodic contextual criteria, we were able to faithfully reproduce the voice and given speaking-style of a speaker and speech corpus [15, 16]. In this paper, we will see how the use of even higher-level selection constraints can make even the prosodic component similarly redundant.

To summarise, the early generations of speech synthesisers were soon able to

reproduce the linguistic content of a message, and the developments described above resulted in an ability to reproduce finer extra-linguistic content; i.e., the speaker-specic characteristics. However, the paralinguistic aspects of speech still remain poorly modelled. Current speech synthesis can function effectively when presenting information by use of a given voice, but it cannot yet perform in a conversational context where laughter, the expression of affect, and the management of discourse ow all take on a greater importance.

3 Expression of Affect

In an effort to produce 'friendlier' speech synthesis, the latest trends in synthesis research have recently become focussed on 'emotion' [17, 18, 19]. The poor take-up of speech technology in general by members of the public is currently attributed, by both the synthesis and recognition communities, to a lack in its ability to process emotion in the speech.

While it may well be true that current speech technology is lacking a 'human' component, is this really best described by the term 'emotion'? I disagree. Or rather, I believe that what many people understand by the wider colloquial application of this term is not well represented by the more limited technical application of the term, as characterised by the 'big-six' emotions of psychological research as illustrated by Ekman and his colleagues [20].

Most speech technology research is now based upon the analysis and modelling of speech databases. These are generally produced under controlled conditions; whether in a recording studio, using the voices of professional speakers to provide 'clean' data, or over the telephone, using the voices of many speakers to collect 'representative' data. The demands of scientic research and of technological developent require balance in the speech data so that they will be representative of the aspects of speech which we wish to reproduce. These controls can take the form of 'phonetic balance', from reading of carefully produced sets of sentences so that each phone is presented in every context of possible use, or of 'sociological balance' so that each sector of the community is 'equally' represented, or of 'content balance' so that all speakers produce a common set of desired utterance types.

The drawback with the above 'scientic' constraints is that we only find what we originally intended to look for. The 'life' is taken out of the data. That is, the data that we produce for research are selected to be representative of those aspects of speech that are generally considered to be important at a given stage of the evolution of the technology, but it is a key point of this paper that they are therefore not representative of the many different ways that ordinary people use speech in the everyday contexts of their social interactions. Data produced as data cannot be as representative of functional interactive speech as that caught in a broader corpus. When confronting researchers with this dilemma, whether in a review of a submitted journal paper or in casual conversation, we often meet the response "Well, what else can we do?". It appears that many of us are aware of the drawbacks of using constructed data but that we nonetheless continue to

follow in the footsteps of our predecessors. Such is the path of scientic research.

So why is this a problem for the processing of emotion in speech? The chain of logic is as follows: (i) emotion is poorly represented in current speech processing, so (ii) emotionally charged speech data should be collected, (iii) the texts must be balanced so that scientic comparisons can be made, so (iv) semanticaly neutral sets of sentences should be produced under various emotions, so (v) actors are recorded producing each sentence in every emotional state, then (vi) perception tests are carried out to 'validate' the data, and (vii) subsequent analyses conrm the clear acoustic characteristics of the different 'emotions'.

This is a very logical progression but it results in a corpus of stereotypical expressions that may have very little to do with how ordinary people vary their speech in actual social interactions. Actors are trained to project what will be readily perceived as a given emotion, and listeners in the perception tests are ofered forcedchoice answers, between alternatives which restrict them from qualifying or elaborating on their 'peceptions' in any way. Furthermore, the 'emotions' that are almost always produced for such data tend to be simple basic ones: sadness, fear, anger, and joy, rather than the more subtle and complex states than result from the interaction of emotions and attitudes arising from interpersonal social interactions. It is rare in everyday life for us to experience or express fear and joy to the extent that they are produced in such 'balanced' data.

Despite the popularity of the keyword 'emotion' in current speech technology research, the question remains as to whether this is in fact the proper direction in which to further our work. Are not 'attitudes' more relevant to spoken interactions? Perhaps we experience boredom or frustration more often that we experience sadness and joy? And show interest more often than we show anger? These more complex expressions of affective states and social relationships are far more common than the expression (or even the experience?) of the basic emotions as illustrated by Ekman in his work on facial expression. Certainly for the use of speech synthesis or recognition in social situations, we need also to be able to reproduce and recognise the more subtle expressions of speaker states and relationships — not just those deliberately produced on demand, but also those which are revealed in spite of a veneer of civilised self-control. Computers need not be able to laugh or cry, but speech synthesis should be able to convey all of the relevant information in speech, and if it is to be used in a conversational context, perhaps in place of people, then it must be as exible and as subtle as the people themselves.

4 A Conversational Corpus

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 [21, 22]. 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, we

persuaded our subjects 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 for us, and a further group were paid to transcribe and annotate this speech data in fine detail. The transcriptions were produced in plain text, using Japanese kana orthography rather than phonetic encoding, but care was taken to transcibe every utterance exactly as it had been spoken, with no effort made to 'cleanup' the transcriptions or correct the grammar.

Transcribers were encouraged to break the speech into the smallest possible utterance chunks by use of a notional 'one-yen-per-line' payment policy. In spite of this, many single 'utterances' included several tens of syllables, often 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 utterances in this corpus were single phrases; 'grunts', or phatic non-verbal speech sounds, made to reassure the listener of the speaker's affective states and discoursal intentions [23, 24]. 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, though the shorter 'grunts' require a new method of treatment for synthesis.

The word 'grunt' carries implications of pre-human or even animal behaviour, but I believe that it is the most appropriate term for the type of phatic communication that takes the place of mutual grooming in human society [25]. As well as the frequent "ummm", "ahhh", "yeah", "uh-uh", etc., I 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 normal. They float to the top of the multigram dictionary [26] by dint of their frequent occurrence, but most can be characterised by the exibility 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.

figure 1 about here

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

On the basis of the above 'interpersonal/infomational' functional distinction, we have categorised the corpus utterances in terms of Itype and A-type functions; the former for the conveyance of *information*, the latter for the expression of *affect* [27, 28]. A framework was proposed (see Figure 1 for an illustration) which describes the twoway giving and getting of I-type and A-type information subject to speaker-state and listener-relationships. For simplicity in a conversational speech synthesis interface, 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: relaxed, formal).

Any given utterance is realised in a discourse subject to the above constraints, and its realisation as speech will therefore vary accordingly. The challenge to synthesisers for conversational speech is to allow the user to specify these constraints simply and easily. In the case of A-type utterances, the framework is more important than the text, which can be relatively freely specified so long as it fulfills the desired social function of the utterance, as we will see below.

5 Functional Unit-Selection

As explained above, we consider there to be two types of utterance in common use in conversational speech; one for transmitting propositional content (I-type) and the other for expressing affect (A-type). While existing speech synthesis technology is arguably quite adequate for the former, the subtlety of prosodic expression and voice-quality (laryngeal phonation settings) required for the latter is beyond the capability of most present systems.

While research is being carried out into signal processing techniques for modifying the voice-source settings, we have yet to find a method that is capable of also matching the sub- and supra-glottal conditions so that a realistic coherent sound can be produced. At present, any modication of the speech signal results in a perceptible degradation which, given that we are trying to control fine modications in vocal setting, such as tenseness and laxness of the voice [29, 30] is unacceptable. The vocal tract can perhaps be adequately modelled as a series of resonant tubes for the purpose of reproducing the basic speech sounds, but for

the fine details of airow required to reproduce the subtle nuances of expression in conversational speech, the model becomes excessively complex.

While not necessarily implying that such a large corpus would be necessary for conversational speech synthesis in different voices or languages, we were able to use the ESP corpus as a test case of what might be possible for concatenative synthesis in the future. Given 5- years of one person's daily conversational speech, we were interested to discover the extent to which the 6th year's speech might be contained within such a corpus.

Our first task was to reduce the data into fundamental units, since segmentation into phone-sized units is no longer necessary, or even desirable, when whole utterances are included in many varied forms, each having different prosodic characteristics, as candidate units. For this we used a form of multigram analysis [26], based on the transcriptions, to determine on statistical grounds the common collocations of frequently-occurring sound sequences in the corpus. This analysis resulted in a dictionary of various-length sequences and a set of probabilities for each so that a subsequent Viterbi process based on the EM algorithm can determine the optimal sequence of segments for any given target utterance.

The multigram analysis provides a speaker-specfiic dictionary of frequently used sound sequences (speech chunks), i.e., a personalised lexicon independent of any linguistic criteria, that models the common speech patterns of the corpus speaker. Frequent phrases and common lexical sequences (e.g., adjective-noun groups and most A-type utterances) tend to be included as intact units with high probabilities in the dictionary, while shorter patterns with even higher probabilities represent the frequent phonetic sequences (or common articulatory gestures) of the speaker. At the lowest level, single phone-sized sounds are also indexed to ensure that any possible sequence of sounds can be generated.

By use of such statistically-determined non-uniform segments for concatenation, whole phrases can be retrieved intact, or constructed from sequences of common articulatory gestures so that a high level of naturalness, retaining the speaker-characteristics, can be maintained in the resulting synthesised speech.

As we saw above, more than half of the utterances can be expected to occur intact, as entire phrases, which can then be further subcategorised according to the prosodic and voice-quality characteristics related to functional differences for the common A-type utterances. With so large a corpus, the task becomes one of selecting the appropriate acoustic realisation of a given phrase rather than that of creating a phrase out of smaller component segments. The original discourse context of the utterance will determine its acoustic characteristics, so rather than code each segment at the lowest parameter levels (which we also do) it is simpler to access the different variants by means of sufcient higher-level contextual features (as illustrated in Figure 1 above).

In parallel with the problem of determining optimal unit size, is the equivalent problem of how to specify such units for input to the synthesiser. Plain text is no longer appropriate when the intention of the speaker is more important than the lexical sequence of the utterance. Instead, we need to enable the user to quickly access a given corpus segment (i.e., a phrase-sized utterance) by means of higherlevel intention-related functional constraints.

figure 2 about here

Figure 2 shows a recent prototype for such a speech synthesis interface. 'Chakai'[2] allows for free input (by typing text into the white box shown at bottom-centre) as well as the 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 enables linking to a conventional CHATR-type synthesiser for creation by unit-selection of I-type utterances not found in the corpus, while providing a fast, three-click, interface for the common A-type utterances which occur most frequently in ordinary conversational speech.

The selection of whole phrases from a large conversational-speech corpus requires specification not just of the function 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. Chakai can be used in almost real-time for conversational interaction. When initiating a topic, typed input is required, and this is presently too slow for real-time use, 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 (e.g., '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 indicated by activating relevant items in a row of smiley-faces (along the top of the figure) from which the user can select the closest to their intended interactional function. No lexical-based selection or keyboard entry is offered, as the function and constraints will determine the text automatically from the suitable candidates available in the corpus for that particular speaker.

The subsequent and main screen (shown in the figure) is for the core part of the conversational interaction. Icons are arranged in four rows, with questions aligned vertically 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. We have tested this interface in actual conversations, and a trained operator can use it in real-time to sustain a conversation for extended periods.

By splitting utterances into three types, we have greatly facilitated the selection process. I-type utterances, being largely unique since they are so content-

[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.

dependent, still have to be laboriously typed in. Frequent phrases which are text-specic can be selected and a choice of speaking styles then offered via the smiley-face icon layer. Grunts, which are the most common type of utterance in casual speech, are the fastest to produce. Each can be generated by simply clicking on the type and its qualier. The corpus has been pre-annotated for the significant parameters of unit-selection so the actual code that produces the segments is very simple (currently 900 lines of perl). And since it is often the case that whole-phrase segments are concatenated, usually with short pauses between them, the naturalness of the resulting speech can be absolute. No further processing is required, thanks to the number and variety of utterances in the corpus, and the multidimensional functional framework that is used for accessing them.

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 satised 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 higherlevel constraints of discourse context and speaker-intention just as the ne 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. Laughter is often produced, and is included in the segments naturally (see online examples at http://feast.his.atr.jp/aesop).

6 Selecting Optimal Phrase-sized Units

It is usually the case that there are very many candidate tokens for any given A-type utterance in the speech corpus. The optimal candidate must be determined according to its prosodic characteristics, but since these units are to be synthesised as pause-delimited whole-phrases, the previous and following contexts become less relevant. The most useful selection criterion is 'interlocutor', which predetermines most of the voice-quality and pitch-range constraints, but we also store and use values of the various prosodic attributes which have been prenormalised by z-score according to utterance type. Selection is biased by the settings of Self and Other levels to produce a candidate more or less 'activated' for synthesis.

A problem in candidate selection is in determining the equivalences between textually different but functionally equivalent utterances. As mentioned above, a morning greeting (for example) can have many different textual representations, and a full 'functional' labelling of the corpus is not yet complete. Similarly, since the speaker-specific colloquial language-use is presumably not known to the user of the synthesiser interface when typing input from the keyboard, if

a choice of phrasing not typically used by the corpus speaker is entered, then no appropriate utterance will be found, even though there may be many functionally equivalent utterances available, each having a slghtly different phrasing. This problem of mapping from the citation forms into the vernacular is being tackled as ongoing work. For this corpus to be of wider use in conversational speech synthesis, a distance-based thesaurus mapping from citation forms to the colloquial usages of the corpus speaker mayl need to be generated.

For extension of this method to different voices, we would need to produce a large corpus of different speaking styles, ideally with different discourse partners being also present during the recordings. Given the experience gained from the rst data collection, and the knowledge that we now have, these recordings would not take another five years but could perhaps be completed within a month. The essential point is to have different interlocutors present during the recordings so that the corpus speaker will be able to naturally adjust her (or his) voice quality and speaking styles and interaction types so that sufcient samples of each type of functional 'grunt' can be obtained naurally. By so constraining the speaker during recordings, we will be able to obtain speech tokens that better represent the speaker's typical daily speech performance, and then to reproduce these highly personal characteristics by concatenation of entire phrases for the A-type utterances that characterise conversational speech.

7 Discussion and Conclusion

This paper has introduced our most recent work on the synthesis of conversational speech, and has shown that the challenges presented by this task are qualitatively different from those of traditional speech synthesis 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 specic prosodic characteristics such as tone-of-voice, realised by differences in laryngeal phonation quality, that can not easily be reproduced by signal processing techniques. The paper has also described our initial attempts to utilise the corpus for concatenative speech synthesis, and has presented a prototype user-interface that allows input acording to speech-act intention, using constraints representing the primary contextual inuences 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 characteristic 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. Samples of the resulting conversational speech synthesis are available on the web at http://feast.atr.jp/laughs. 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 and offers them as one form of its solution.

Acknowledgements

This work is partly supported by the Japan Science & Technology Corporation (JST), partly by the National Institute of Information and Communications Technology (NiCT), and partly by the Ministry of Public Management, Home Affairs, Posts and Telecommunications, Japan. The author is grateful to the management of ATR, especially to the Department of Emergent Communications in the Network Informatics lab and to the Media Information Science Laboratory (MIS) for their continuing encouragement and support.

References

[1] Holmes, J.N., Mattingley, I.G. & Shearme, J.N., Speech synthesis by rule, Language and Speech 7, 127-143, 1964.

[2] Mattingly, I.G., Experimental methods for speech synthesis by rules, IEEE Trans. AU 16, 198-202, 1968.

[3] Fant, G. Acoustic Analysis and Synthesis of Speech with Applications to Swedish, Ericsson Technics 15, 3-108, 1959.

[4] Carlson, R. & Granstrom, B., A text-to-speech system based entirely on rules, Proc. IEEE-ICASSP76, 686-688, 1976.

[5] Allen, J., Hunnicutt, M. S. & Klatt, D.H., From text to speech. The MITalk system, Cambridge University Press, Cambridge UK, 1987.

[6] Klatt, D.H., The Klattalk text-to-speech conversion system, Proc. IEEE-ICASSP82, 1589-1592, 1982.

[7] Olive, J.P., Rule synthesis of speech from dyadic units, Proc. IEEE-ICASSP77, 568-570, 1977.

[8] Olive, J.P. 1980, A scheme for concatenating units for speech synthesis, Proc. IEEE-ICASSP80, 568-571.

[9] Olive, J.P. & Liberman, M., A set of concatenative units for speech synthesis, In: J.J. Wolff and D.H. Klatt Eds., ASA*50 Speech Communication Papers, 515-518, 1979.

[10] Sagisaka, Y., Speech synthesis by rule using an optimal selection of nonuniform synthesis units, Proc. IEEE-ICASSP88, 679-682, 1988.

[11] Church, K., Stress assignment in letter to sound rules for speech synthesis. In ACL Proceedings, 23rd Annual Meeting, pages 246253, Morristown, NJ, 1985. Association for Computational Linguistics.1985.

[12] Campbell, W.N. & Wightman, C.W. 1992, Prosodic coding of syntactic structure in English speech, Proc. ICSLP92, Banff, Canada, 1167-1170.

[13] Campbell, W.N., Synthesis units for natural English speech, Transactions of the Institute of Electronics, Information and Communication Eng, Vol. SP 91-129, 55-62, 1992.

[14] Campbell, W.N., CHATR: A High-Denition Speech Re- Sequencing System, proc Eurospeech'95, Madrid/Spain, 1995.

[15] Campbell, W. N. and Black, A. W. CHATR a multi-lingual speech re-sequencing synthesis system. Technical Report of IEICE SP96-7, 45-52, 1996.

[16] CHATR Speech Synthesis: http://feast.his.atr.jp/chatr

[17] Iida, A., Campbell, N. and Yasumura, M. Design and Evaluation of Synthesised Speech with Emotion. Journal of Information Processing Society of Japan Vol. 40, 1998.

[18] Iida, A., Campbell, N., Iga, S., Higuchi, Y,. and Yasumura, Y., A speech synthesis system with emotion for assisting communication . In Proceedings of the ISCA Workshop on Speech and Emotion, pages 167-172, Belfast, 2000.

[19] Schroder, M., et. al., Acoustic correlates of emotion dimensions in view of speech synthesis, pp.87-90, In Proc Eurospeech 2001, Denmark, 2001.

[20] Ekman, P., Universals and cuntural differences in facial expression of emotion, in J., K. Cole (Eds), Nebraska Symposium on Motivation, pp.207-282, Lincoln, University of Nebrasaka Press, 1972.

[21] 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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

[28] 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 Articial Intelligence Series, 2004.

[29] Campbell, N., & Mokhtari, P., Voice quality: the 4th prosodic dimension, pp.2417-2420 in Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'03), Barcelona, Spain, 2003.

[30] Campbell, N., Perception of Affect in Speech - towards an Automatic Processing of Paralinguistic Information in Spoken Conversation , in Proc ICSLP 2004.

Figures

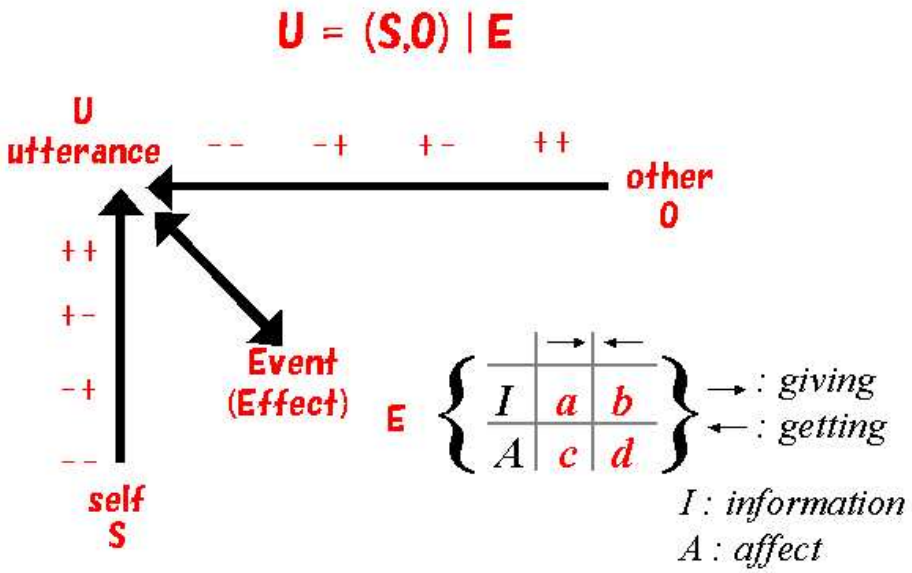

Figure 1: A framework for specifying the characteristics of an utterance according to speaker-state, relationship with the listener, and speech-act type.

Figure 2: 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 availability in the corpus, from which an utterance having the appropriate speech characteristics 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.