Conversational Speech Synthesis and the Need for Some Laughter

Nick Campbell

Abstract—This paper reports progress in 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 recognizable 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 modification 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 modeled by rule, but they are also expresssed through nonverbal 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.

Index Terms—Affect, conversation, emotion, expression, laughter, nonverbal, social interaction, speech synthesis.

I. 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, with the notable exception of modern Conversation and Discourse Analysis [3] and projects like Talkbank,[1] it has been treated either as a "system of sounds" or as a "media-transformed" version of text, to be analyzed in written form through the use of transcriptions. This is understandable, since the technology for recording and analyzing oral interactions has until recently been both expensive and lacking in portability. As a result, "speech" is not yet well understood from the standpoint of "communication."

We now 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. However, 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*, and 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.

The emphasis in speech data collection has been on maximizing speaker numbers in order to produce speaker-independent models, rather than on modeling 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. 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, perhaps on the assumption that meaning can be understood from linguistic context alone.

Similarly for speech synthesis research, based on the early notion of synthesizers 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 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 nonverbal 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 [4]. This paper tackles the issue of how to synthesise such nonverbal phatic utterances for use in conversational speech.

II. Corpus-Based Speech Synthesis

Looking back across the long history of speech synthesis research, we can see in retrospect a clear evolution from the

Manuscript received November 9, 2005; revised March 29, 2006. This work was supported in part by the Japan Science and Technology Corporation (JST), in part by the National Institute of Information and Communication Technology (NiCT), and in part by the Ministry of Public Management, Home Affairs, Posts, and Telecommunications, Japan.

The author is with the National Institute of Information and Communication Technology and the Speech and Acoustic Processing Department, ATR Spoken Language Communication Laboratory, Keihanna Science City, Kyoto 619-0288, Japan (e-mail: nick@atr.jp; nick@nict.go.jp).

Digital Object Identifier 10.1109/TASL.2006.876131

[1] [Online]. Available: http://www.talkbank.org.

modeling of phonetic states to the modeling of utterance characteristics. The pioneering work of Fant in Sweden [5], [6] and Klatt and his colleagues in the U.S. [7], [8] was founded on a phonetic view of speech as a sequence of phones, modulated by prosody to represent syntactic and semantic content. Olive [9]–[11], 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 centers. Sagisaka in Japan [12] extended this paradigm shift by concatenating nonuniform 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 centers or steady states. The variation itself encodes information in 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 [13], and Campbell's work extended the aforementioned trend by incorporating prosodic contexts among the selection criteria of units for concatenation from a speech corpus [14], [15]. Although a small step in terms of unit-selection, this rendered much of the signal-processing unneccessary [16] and enabled concatenation of the speech segments intact, without the need for potentially damaging signal modification. By simply joining phone-sized segments which had been selected according to both phonetic and prosodic contextual criteria, concatenative synthesis was able to faithfully reproduce the voice and given speaking-style of a speaker and speech corpus[2] [17], [18]. In this paper, we will see how the use of even higher-level selection constraints can make even the prosodic component similarly unneccessary.

To summarize, 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-specific characteristics. However, the paralinguistic aspects of speech still remain poorly modeled. 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 now all take on a greater importance.

III. Expression of Attitude and Affect

The slow take-up of speech synthesis by the public is currently attributed by many members of the speech processing community to a lack in its ability to express emotion in the speech. In an effort to produce "friendlier" speech synthesis, the latest trends in synthesis research have, therefore, become focussed on "emotion" [21]–[25].

However, what many understand by the wider colloquial application of this term is perhaps not well represented by the more limited technical application of the term, usually characterized by the "big-six" emotions of psychological research as illustrated by Ekman and his colleagues [26].

It may well be true that current speech technology is lacking a "human" component, but is this really best described by the term "emotion"? I disagree. Or rather, I believe that the link between this level of human interaction and what takes place in conversational spoken interactions is slim, or at too fundamental a level to be directly modeled in speech technology.

There is of course a wealth of relevant and interesting research into human emotions (see especially the recent work of the Humaine Group[3] in Europe, the Affective Computing Groups in the U.S.,[4] and the Kansei-related developments in the Far East[5]), but few of these (with some notable exceptions, e.g., [27]) directly address the need to express attitudes and relationships through use of the voice in spoken interactions. They focus more on recognizing and synthesizing the inner feelings and states of a person than on modeling the interactive processes of a spoken interaction.

Perhaps what needs to be modeled first in conversational speech synthesis is closer to what Malinowski [28] termed "phatic communion," i.e., "a type of speech in which the ties of union are created by a mere exchange of words," what Jakobsen describes as having the function of "maintaining an open channel of communication between interlocutors" [29], and Sherer includes under "interpersonal stances" [30].

Most speech technology research is now based upon the analysis and modeling 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 scientific 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 "scientific" 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.

So why is this a problem for the processing of emotion in speech? The chain of logic is as follows. 1) Emotion is poorly

[2] CHATR Speech Synthesis. [Online]. Available: http://feast.atr.jp/chatr.

[3] Humaine Emotion Research. [Online]. Available: http://emotion-research.net

[4] Affective Computing. [Online]. Available: http://affect.media.mit.edu.

[5] Kansei Engineering. [Online]. Available: http://www.ergolabs.com/kansei_engineering.htm.

represented in current speech processing, so 2) emotionally charged speech data should be collected, 3) the texts must be balanced so that scientific comparisons can be made, so 4) semanticaly neutral sets of sentences should be produced under various emotions, so 5) actors are recorded producing each sentence in every emotional state, then 6) perception tests are carried out to "validate" the data, and 7) subsequent analyses confirm 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 offered forced-choice 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 recognize 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 civilized 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 flexible and as subtle as the people themselves.

IV. A Conversational Corpus

In order to discover what the more likely distributions of affective or emotional expressions might be, we collected a corpus of everyday conversational speech, which has been reported in detail elsewhere (the ESP corpus[6] [31], [32]). 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.

TABLE I
Three Levels of Labelling for Describing Each Utterance, Including Use of Six-Level Forced-Choice Tendency Scales

level 1	STATE (about the speaker)
purpose	a discourse-act/DA label
emotion	happy/sad/angry/calm
mood	worried/tense/frustrated/troubled/...
interest	a 6-point scale from +3 to -3, omitting 0
confidence	a 6-point scale from +3 to -3, omitting 0
level 2	STYLE (about the speech)
type	speaking-style label (open-class)
purpose	a discourse-act label (closed-class)
sincerity	insisting/telling/feeling/recalling/acting/...
manner	polite/rude/casual/blunt/sloppy/childish/sexy/...
mood	happy/sad/confident/diffident/soft/aggressive/...
bias	friendly/warm/jealous/sarcastic/flattering/aloof/...
level 3	VOICE (about the sound)
energy	a 6-point scale from +3 to -3, omitting 0
tension	a 6-point scale from +3 to -3, omitting 0
brightness	a 6-point scale from +3 to -3, omitting 0
level 0	labeller
confidence	a 6-point scale from +3 to -3, omitting 0

6-point values:	negative	positive
'very noticeable'	-3	3
'noticeable'	-2	2
'only slightly noticeable'	-1	1

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, 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 to 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. This yielded a total of about 450 000 utterances for one prolific speaker. Within these, more than 200 000 utterances could be classed as "grunts" of which the most frequent ($n = 4$) occurred more than 10 000 times each and 15 types occurred more than 1000 times each. The number of unique utterances from this speaker was 214 378.

All utterances were transcribed and tagged for interlocutor, and a subset representing slightly more than 10 percent of the corpus was labeled by hand for discourse act and speaking-style features (see Table I) paying particular attention to what Chafe terms "regulatory intonation units" [3]. We labeled speaker state features independently of speech style (the former requiring longer-time windows than the latter which can be performed utterance by utterance) and attempted to characterize the voice quality features of each utterance. We then trained decision trees to perform automatic labeling of the remaining utterances in the corpus using the following features: utterance duration, f0-range, f0-variation, f0-maximum, f0-minimum, f0-mean, position of f0 peak in the utterance, position of f0-minimum in the utterance, power-range, power-variation, power-maximum, power-minimum, power-mean, position of power-peak in the utterance, position of the power-minimum in the utterance. The tree correctly predicted 68% (or 217/315) of speaker-state categories using 26 leaf nodes.

[6] The Expressive Speech Processing Project Web Pages. [Online]. Available: http://feast.atr.jp

The majority of utterances in this corpus were single phrases: "grunts," or phatic nonverbal speech sounds, made to reassure the listener of the speaker's affective states and discoursal intentions [33], [34]. As well as the frequent "ummm," "ahhh," "yeah," "uh-uh," etc., (or their Japanese equivalents), 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 [35] by dint of their frequent occurrence, but most can be characterized by the flexibility and variety of their prosody. Very few can be adequately interpreted from the plain text alone. Laughs were very frequent, as were back-channel utterances and fillers,[7] 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.

On the basis of the above "interpersonal/informational" functional distinction, we have categorized 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* [36]. A framework has been proposed which describes the two-way giving and getting of I-type and A-type information subject to speaker–state and listener-relationships.

We recognize that an utterance has a direct relationship to a discourse event, and that the relationship is subject to constraints from two dimensions related to the factors; i.e., 1) influences from the speaker's own states (Self), and 2) influences from the listener and the discourse context (Other). The relationship is two-way so the framework can be used for interpreting a spoken utterance to retrieve its intended discourse "effect," or for generating an utterance from a discourse "event."

For simplicity in the current conversational speech synthesis interface, we have assumed four levels of each constraint, representing high or low activation of "mood" and "interest," in conjunction with "close" or "distant" listener relations:

- 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 "energized" 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 realized in a discourse subject to the above constraints, and its realization as speech will, therefore, vary accordingly. The challenge to synthesizers 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 next.

V. Functional Unit-Selection

As explained previously, we consider there to be two types of utterances 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 modification of the speech signal results in a perceptible degradation which, given that we are trying to control fine modifications in vocal setting, such as tenseness and laxness of the voice [37], [38] is unacceptable. The vocal tract can perhaps be adequately modeled as a series of resonant tubes for the purpose of reproducing the basic speech sounds, but for the fine details of airflow 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 five years of one person's daily conversational speech, we were interested to discover the extent to which the sixth 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 [35], 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 determines the optimal sequence of segments for any given target utterance.

The multigram analysis provides a speaker-specific dictionary of frequently used sound sequences (speech chunks), i.e., a personalized 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.

[7] 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 nonlinguistic (affective) communication can occur.

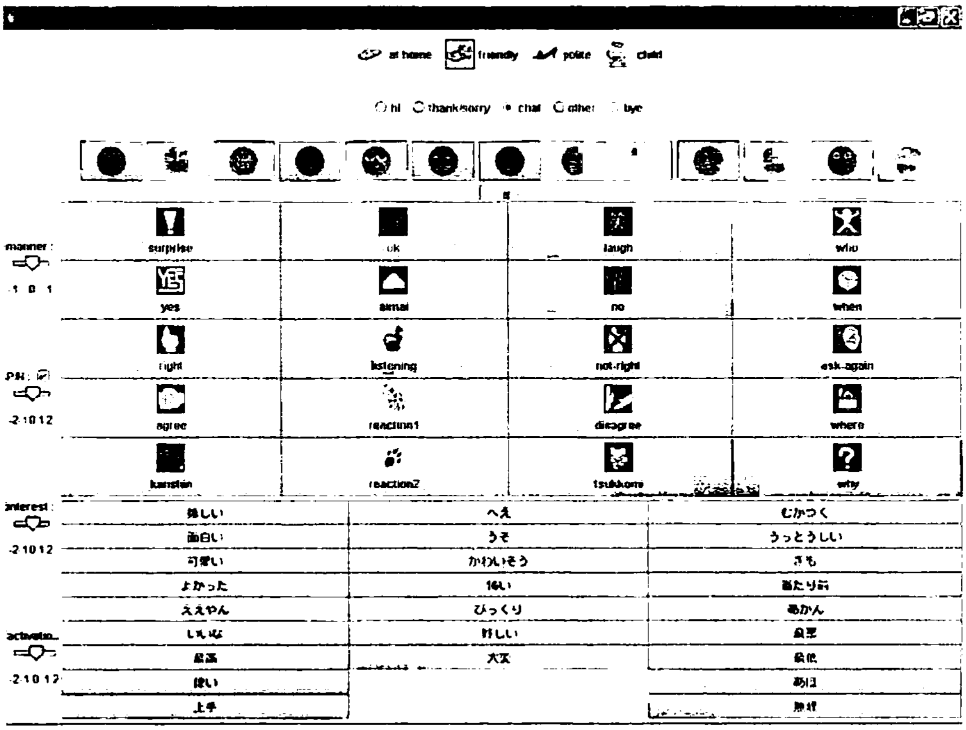

Fig. 1. Chakai Conversational Speech Synthesis interface. By clicking on a discourse-act icon, a choice of emoticons is displayed in the upper section of the screen, according to availability in the corpus, from which an utterance having the appropriate speech characteristics can be selected. Utterances 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.

By use of such statistically determined nonuniform 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 synthesized 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 subcategorized 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 realization 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 sufficient higher-level contextual features.

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 higher level intention-related functional constraints.

Fig. 1 shows a recent prototype for such a speech synthesis interface. "Chakai"[8] allows for free input (by typing text into the white box shown at bottom-center) as well as the fast selection of various frequently-used phrases and, in addition, an icon-based discourse-act selection facility for the most common types of "grunt." This format enables linking to a conventional CHATR-type synthesizer 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. A demonstration version of this software (with limited functionality) can be downloaded from http://www.feast.atr.jp/chakai for evaluation of the interface.

The sliders on the left represent 1) manner, 2) politeness, 3) interest, and 4) activation, providing additional bias in the selection of the phrases. The granularity of the settings (three or five states each) reflects the labels used in the manual annotation of the training data. A subset representing slightly more than 10% of the corpus was labeled by hand for these four categories and the remaining data was classified using support vector machines trained on the acoustic characteristics of the labeled data. Ten-fold cross-validation indicated varying degrees of classification accuracy (ranging from 74% for activation to 48% for politeness).

Chakai can be used in almost real-time for conversational interaction. 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. When initiating a topic, typed input is required, and this is presently too slow for

[8] 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.

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, subcategorized 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 short conversation of about five minutes.

By splitting utterances into three types, we have greatly facilitated the selection process. I-type utterances, being largely unique since they are so content-dependent, still have to be laboriously typed in. Frequent phrases which are text-specific 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 then selecting from among its smiley-face qualifiers. The corpus has been preannotated 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.

In order to search the corpus for the most appropriate candidate utterances which have a meaning close to the intended utterance, we use the public-domain, open-source, web-browser search-engine software SWISH-E.[9] This has been incorporated as part of the front-end for selecting utterances from the speech corpus. The search engine provides a Google-like index of all utterances in the corpus and enables refined retrieval of selected utterance samples by the use of AND and NOT conditions in the search key.

For example, searching a database of 436 961 utterances from one speaker, using "mama AND papa NOT X" [i.e., look for all sentences containing both the word "mama" and the word "papa" but exclude all those that contain the symbol "X" (which indicates a noisy recording)] as the search-key, yielded four results, with a run time of 0.077 s and a search time of 0.065 s on a notebook PC. A search for "honma NOT X" took 0.094 s (search-time 0.083 s) and yielded 2498 results. Given these high-speed access facilities, the remaining problem is to produce an efficient language-model so that a closer approximation to the idiomatic and colloquial nature of the corpus contents can be produced automatically. This largely remains as future 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 realized, but we are satised that the framework 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. Laughter is often produced, and is included in the segments naturally (see online examples at http://www.feast.atr.jp/aesop).

VI. Selecting Phrase-Sized Units

In conventional unit-selection, phone-sized units are selected from a database according to prosodic (i.e., the target cost) and spectral characteristics (i.e., the join cost) to ensure smooth and natural-sounding speech output. However, when the unit to be synthesized is an entire phrase, typically surrounded by pauses in the speech, then the continuity constraints (i.e., spectral smoothness) cease to be relevant and the optimal candidate can be selected purely on the basis of its prosodic characteristics.

According to a linear discriminant analysis, the single most influential factor in determining the prosodic and speaking-style characteristics of an A-type utterance is the "Class of Interlocutor." All corpus utterances are tagged for interlocutor ID, and these tags were subsequently grouped manually into a smaller number of classes (e.g., "close friend," "distant friend," "family member," "small child," etc.) so that the setting of the "Other" parameter can be used directly as a criterion in phrase selection.

The second most influential selection parameter is "Activation," a measure related to speech-rate, pitch excursion, and energy in the speech. This is heuristically mapped to the setting of the "Self" parameter by a set of rules sensitive to spectral slope, fundamental-frequency, and signal energy so that a high setting results in brighter-sounding phrases being selected from the corpus. All utterances are acoustically analyzed for prosodic and phonation characteristics, and 28 z-score normalized values are ranked to produce a list of candidate phrases in the selection process.

[9] SWISH-E—Simple Web Indexing System for Humans, Enhanced Version. [Online]. Available: http://swish-e.org.

A remaining problem in candidate selection is in determining the equivalences between textually different but functionally equivalent utterances. As mentioned previously, a morning greeting (for example) can have many different textual representations, and a full "functional" labeling of the corpus is not yet complete. Similarly, since the speaker-specific colloquial language-use is presumably not known to the user of the synthesizer 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 slightly 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 may need to be generated.

VII. Discussion and Conclusion

This paper has introduced our most recent work on the synthesis of conversational speech. It limits itself to one particular conversational aspect of interactive speech—the short affect-burst—and presents a novel method for incorporating these into conventional concatenative speech synthesis.

It 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 specific prosodic characteristics such as tone-of-voice, realized by differences in laryngeal phonation quality, that cannot easily be reproduced by signal processing techniques. The paper has also described our initial attempts to utilize the corpus for concatenative speech synthesis, and has presented a prototype user-interface that allows input according to discourse-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 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 first 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 sufficient 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 characterize conversational speech.

We currently use hardware voice modification (passing the output through a Roland UA-100 Voice Transformer) to disguise the original speaker's voice and reduce the spectral detail while preserving the prosodic details in the speech. This allows adjustment of the formants and pitch to simulate a male, female, or child speaker from the original female waveforms. However, even after voice conversion, the persona projected by the conversational idiosyncracies of a given speaker may sometimes be at odds with what the user of the system wants to convey.

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://www.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.

Acknowledgment

The author would like to thank the management of ATR, especially to the Department of Emergent Communications in the Network Informatics Laboratory and to the Media Information Science Laboratory (MIS) for their continuing encouragement and support. The author would also like to thank the anonymous reviewers for their helpful comments, whose suggestions made this manuscript far better than the originally submitted version.

References

[1] J. N. Holmes, I. G. Mattingley, and J. N. Shearme, "Speech synthesis by rule," *Lang. Speech*, vol. 7, pp. 127–143, 1964.

[2] I. G. Mattingly, "Experimental methods for speech synthesis by rules," *IEEE Trans. Audio Electroacoust.*, vol. 16, no. 2, pp. 198–202, Jun. 1968.

[3] W. L. Chafe, "Prosodic and functional units of language," in *Talking Data: Transcription and Coding in Discourse Research*, J. A. Edwards and M. D. Lampert, Eds. Hillsdale, NJ: Lawrence Erlbaum, 1993, pp. 33–43.

[4] N. Campbell and D. Erickson, "What do people hear? a study of the perception of nonverbal affective information in conversational speech," *J. Phonetic Soc. Jpn.*, vol. 7, no. 4, pp. 9–28, 2004.

[5] G. Fant, "Acoustic analysis and synthesis of speech with applications to swedish," *Ericsson Technics*, vol. 15, pp. 3–108, 1959.

[6] R. Carlson and B. Granstrom, "A text-to-speech system based entirely on rules," in *Proc. IEEE ICASSP*, 1976, pp. 686–688.

[7] J. Allen, M. S. Hunnicutt, and D. H. Klatt, "From text to speech," in *The MITalk System*. Cambridge, U.K.: Cambridge Univ. Press, 1987.

[8] D. H. Klatt, "The Klattalk text-to-speech conversion system," in *Proc. IEEE ICASSP*, 1982, pp. 1589–1592.

[9] J. P. Olive, "Rule synthesis of speech from dyadic units," in *Proc. IEEE ICASSP*, 1977, pp. 568–570.

[10] J. P. Olive, "A scheme for concatenating units for speech synthesis," in *Proc. IEEE ICASSP*, 1980, pp. 568–571.

[11] J. P. Olive and M. Liberman, "A set of concatenative units for speech synthesis," in *ASA' 50 Speech Communication Papers*, J. J. Wolff and D. H. Klatt, Eds., 1979, pp. 515–518.

[12] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of nonuniform synthesis units," in *Proc. IEEE ICASSP*, 1988, pp. 679–682.

[13] K. Church, "Stress assignment in letter to sound rules for speech synthesis," in *Proc. ACL 23rd Annu. Meeting*, Morristown, NJ, 1988, pp. 2426–2426.

[14] W. N. Campbell and C. W. Wightman, "Prosodic coding of syntactic structure in English speech," in *Proc. ICSLP*, Banff, AB, Canada, 1992, pp. 1167–1170.

[15] W. N. Campbell, "Synthesis units for natural English speech," *Trans. Inst. Electron., Inf. Commun. Eng.*, vol. SP 91-129, pp. 55–62, 1992.

[16] A. W. Black and W. N. Campbell, "Optimising selection of units from speech databases for concatenative synthesis," in *Proc. Eurospeech*, Madrid, Spain, 1995, pp. 81–84.

[17] C. Min, H. Peng, H. Y. Yang, and E. Chang, "Selecting nonuniform units from a very large corpus for concatenative speech synthesizer," in *Proc. ICASSP*, Salt Lake City, UT, 2001, pp. 785–788.

[18] A. K. Syrdal, "Phonetic effects on listener detection of vowel concatenation," in *Proc. Eurospeech*, 2001, pp. 979–982.

[19] E. Klabbers, K. Stober, R. Veldhuis, P. Wagner, and S. Breuer, "Speech synthesis development made easy: The Bonn open synthesis system," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 521–524.

[20] Y. Stylianou and A. K. Syrdal, "Perceptual and objective detection of discontinuities in concatenative speech synthesis," in *Proc. ICASSP*, 2001, pp. 837–840.

[21] R. Cowie, E. Douglas-Cowie, and C. Cox, "Beyond emotion archetypes: Databases for emotion modeling using neural networks," *Neural Netw.*, vol. 18, pp. 371–388, 2005.

[22] J. Trouvain and M. Schroeder, "How (not) to add laughter to synthetic speech," in *Proc. Workshop Affective Dialogue Syst.*, Kloster Irsee, Germany, 2004, pp. 229–232.

[23] A. Iida, N. Campbell, S. Iga, Y. Higuchi, and Y. Yasumura, "A speech synthesis system with emotion for assisting communication," in *Proc. ISCA Workshop Speech Emotion*, Belfast, U.K., 2000, pp. 167–172.

[24] M. Schroeder *et al.*, "Acoustic correlates of emotion dimensions in view of speech synthesis," in *Proc. Eurospeech*, Aalborg, Denmark, 2001, pp. 87–90.

[25] ——, "Dimensional emotion representation as a basis for speech synthesis with nonextreme emotions," in *Proc. Workshop Affective Dialogue Syst.: Lecture Notes in Comput. Sci.*, Kloster Irsee, Germany, 2004, pp. 209–220.

[26] P. Ekman, "Universals and cultural differences in facial expression of emotion," in *Nebraska Symposium on Motivation*, J. K. Cole, Ed. Lincoln, NB, 1972, pp. 207–282.

[27] O. Turk, M. Schroeder, B. Bozkurt, and L. M Arslan, "Voice quality interpolation for emotional text-to-speech synthesis," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 797–800.

[28] B. Malinowski, *The Problem of Meaning in Primitive Languages*. London, U.K.: Routledge and Kegan Paul, 1923, pp. 146–152. Supplement to C. Ogden and I. Richards, *The Meaning of Meaning*.

[29] R. Jakobson, "Linguistics and poetics," in *Style in Language*, T. A. Sebeok, Ed. Cambridge, MA: MIT Press, 1960, pp. 350–377.

[30] K. R. Scherer, "Psychological models of emotion," in *The Neuropsychology of Emotion*, J. Borod, Ed. Oxford, U.K.: Oxford Univ. Press, 2000, pp. 137–162.

[31] N. Campbell, "Recording techniques for capturing natural everyday speech," in *Proc. Language Resources Evaluation Conf. (LREC)*, Las Palmas, Spain, 2002, pp. 2029–2032.

[32] ——, "Speech and expression; the value of a longitudinal corpus," in *Proc. Language Resources and Evaluation Conf. (LREC)*, Lisbon, Portugal, 2004, pp. 183–186.

[33] ——, "Getting to the heart of the matter; speech as expression of affect rather than just text or language," *Language Resources and Evaluation*, vol. 39, no. 1, pp. 109–118, 2005.

[34] N. Campbell, "Specifying affect and emotion for expressive speech synthesis," in *Computational Linguistics and Intelligent Text Processing, Proc. CICLing-2004, Lecture Notes in Computer Science*, A. Gelbukh, Ed. New York: Springer-Verlag, 2004.

[35] S. Deligne and F. Bimbot, "Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams," in *Proc. ICASSP*, 1995, pp. 169–172.

[36] N. Campbell, "Extra-semantic protocols; input requirements for the synthesis of dialogue speech," in *Affective Dialogue Systems, Springer Lecture Notes in Artificial Intelligence Series*, E. Andre, L. Dybkjaer, W. Minker, and P. Heisterkamp, Eds. New York: Springer-Verlag, 2004, pp. 221–228.

[37] P. Alku and E. Vilkman, "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering," *Speech Commun.*, vol. 18, no. 2, pp. 131–138, 1996.

[38] N. Campbell, "Perception of affect in speech—toward an automatic processing of paralinguistic information in spoken conversation," in *Proc. ICSLP*, 2004, pp. 881–884.

Nick Campbell received the Ph.D. degree in Experimental psychology from the University of Sussex, Sussex, U.K.

He is currently engaged as a Chief Researcher in the Department of Acoustics and Speech Research, Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan, where he also serves as Research Director for the JST/CREST Expressive Speech Processing and the SCOPE "Robot's Ears" projects. He was first invited as a Research Fellow at the IBM U.K. Scientific Centre, where he developed algorithms for speech synthesis, and later at the AT&T Bell Laboratories, where he worked on the synthesis of Japanese. He served as Senior Linguist at the Edinburgh University Centre for Speech Technology Research before joining ATR in 1990. His research interests are based on large speech databases, and include nonverbal speech processing, concatenative speech synthesis, and prosodic information modeling. He spends his spare time working with postgraduate students as Visiting Professor at the Nara Institute of Science and Technology (NAIST), Nara, Japan, and at Kobe University, Kobe, Japan.