IEICE TRANS. INF. & SYST., VOL.E88–D, NO.3 MARCH 2005

INVITED PAPER Special Section on Corpus-Based Speech Technologies

Developments in Corpus-Based Speech Synthesis: Approaching Natural Conversational Speech

Nick CAMPBELL[†a)], *Nonmember*

SUMMARY This paper describes the special demands of conversational speech in the context of corpus-based speech synthesis. The author proposed the CHATR system of prosody-based unit-selection for concatenative waveform synthesis seven years ago, and now extends this work to incorporate the results of an analysis of five-years of recordings of spontaneous conversational speeech in a wide range of actual daily-life situations. The paper proposes that the expresion of affect (often translated as 'kansei' in Japanese) is the main factor differentiating laboratory speech from real-world conversational speech, and presents a framework for the specification of affect through differences in speaking style and voice quality. Having an enormous corpus of speech samples available for concatenation allows the selection of complete phrase-sized utterance segments, and changes the focus of unit selection from segmental or phonetic continuity to one of prosodic and discoursal appropriateness instead. Samples of the resulting large-corpus-based synthesis can be heard at http://feast.his.atr.jp/AESOP.

key words: *speech synthesis, corpora, concatenation, paralinguistic information, communication, affect*

1. Introduction

In classical Newtonian science, we assume *ceteris paribus*, that all other things being equal, we can alter one variable and examine the effects of its change on all the other so-called dependent variables. In much of the history of speech science, this has been the paradigm of choice, but this view ignores the multiplicity of interdependencies between the variables and reduces their interactions to a mere mechanical process. However, speech is a uniquely human means of comunication that has evolved over the centuries to incorporate delicacies of not just linguistic, but also discoursal and interpersonal factors for the simultaneous communication of both proposition and affect. Like global weather systems, it is a massively complex system of interactions, but unlike weather, each combination of variables can be interpreted as displaying a complex meaning.

Recent advances in speech research have benefitted greatly from the availablility of large corpora, rather than having to rely on laboratory samples, and with the rapid developments in computing facilities and the ready availablility of statistical software for analysis and modelling, we have begun the explore the interactions *in vivo*. The new paradigm alllows us to examine speech in situ, and to model its relevant variables not in isolation but in interaction. Of course, the amounts of data required for such research are orders of magnitude greater than those required for Newtonian analyses, but since the tools and the methods are now available, it is only a matter of time before the corpora are produced and a new generation of models can be evaluated.

The speech signal is as much influenced by its external environment as it is by its liguistic variables, and only by corpus-based research can we begin to model the overall effects of a given speech utterance. Nowhere is this more important than in speech synthesis, where the machine produces sounds that are to be interpreted by a human (through both ear and brain) as speech. Our long experience of interpreting speech sounds (which being probably pre-natal predates our experiences of vision, touch, and smell) has taught us to interpret each and every variation as meaningful.

The current sparcity of paralinguistic and extralinguitic information, and the lack of a model of their complex interactions results in computer synthesised speech sounding mechanical and "lacking in emotion". There is considerable research effort currently being concentrated into the synthesis (and recognition) of emotional speech [1]–[7] but this paper will argue that "emotion" is the wrong term. Based upon our analyses of a very large corpus of spontaneous and ecologically well-situated conversational speech, we propose a more complex but elementary framework for the specification of affect in interactive speech.

2. The Phonetic View of Speech

In the phonetic view, speech is seen as made up of a sequence of basic sounds that are modulated by prosody to portray a given meaning. It is generally assumed that the complex signal can be decomposed into a sequence of basic components which can be sufficiently transcribed as phonemes. The prosodic modulation is presumed not to have a direct effect on the phonemes themselves, but to add an extra-segmental layer of information which indicates how they should be grouped and interpreted [8]. Allophonic variation is seen as predictable and derived from contextual influences that can be predicted from the phone sequence.

In this view, the linguistic framework of an utterance is generally considered as defining all the relevant information concerning its meaning and speech production characteristics. Any extra-linguistic (speaker age, sex, size, health, etc) or paralinguistic (speaker state, affect, emotion, intention, etc) information is considered to be incidental. An utterance is considered to be well described by a narrow phonetic transcription alone, and no prosodic information is required

Manuscript received July 2, 2004.

[†]The author is with the Department of Emergent Communication of the ATR Network Informatics Laboratories, Kyoto-fu, 619–0288 Japan.

a) E-mail: nick@atr.jp

DOI: 10.1093/ietisy/e88–d.3.376

except for the special cases of accent marking, focal stress and phrasal boundaries. Such a transcription can be substituted by its orthographic text equivalent with almost no loss of essential information.

This phonetic view of speech has guided most of the developments in speech synthesis research (e.g., [10]–[20]). D. H. Klatt, for example, one of the early pioneers of computer speech synthesis [21] produced tables of parameters for the specification of each phoneme, with rules for their interpolation and prosodic modification. His input was plain text, and his research was carried out with the aim of developing a "reading-machine". The prosodic modifications of the phonemic sequence served to highlight phrasal boundaries and semantic focus, by raising and lowering the pitch, and lengthening or shortening the segment durations accordingly. Little attention was given to voice quality issues except to mark a change of sex or speaker.

3. The Affective View of Speech

In an affect-based view of speech communication [22]–[24], the linguistic component takes a subsidiary place to the more social aspects of communication that are revealed by the voice and speaking-style parameters [25]. Here, the prosodic overlay begins to take on a more significant meaning. Rather than just signal or reinforce syntactic or semantic relations that could perhaps be derived from the text alone (which is how most speech synthesiser front-ends predict them in the first place [26]), prosody signals an affective layer of communicative information that is superimposed on the linguistic (or non-verbal) content.

In previous work [27], [28] we have proposed that any given conversational speech utterance can be categorised into either I-Type or A-Type classes, where I-Type indicates a predominance of propositional content, and A-Type indicates a predominance of affect in the utterance. I-Type utterances can be safely characterised by a transcription of their linguistic content alone. A-Type utterances, however, can not be adequately understood from just their linguistic content, and require in addition a prosodic specification to indicate the speaker-state and speaker-listener relationships as displayed through significant variations in the speaking-styles and voice qualities.

Being social entities, we humans do not talk just to convey information, but also to form social bonds, to display short-term and long-term relationships, and to simply enjoy (or not) being together. Laughter, for example, is a common feature of interactive speech, but is not yet modelled in synthesis. All such interpersonal and discourse-related information can be reliably conveyed by a sound recording of an utterance and so must be physically represented in the speech signal as well as controlled by the speaker to be perceived by the listener. From this point of view, the linguistic component of the speech assumes a distinctly lower degree of importance, and prosodic information relating to speaking styles and affective display comes more into the foreground.

4. Describing Conversational Acts

This work is based upon an analysis of the JST/CREST ESP Corpus which has been described in detail elsewhere (see for example [29], [30] and related works). It consists (partly) of long-term high-quality recordings of daily-life conversations in which informants wore head-mounted microphones and recorded their spoken interactions with a variety of interlocutors throughout the day for a period of almost five years.

We now maintain that for the specification of a conversational utterance (i.e., a speech event) for concatenative synthesis from such a very large corpus, we first need to determine both the directionality and the function of the event; i.e., whether that utterance is intended primarily to convey or to elicit information (I-Type), or to display or elicit-display-of affect (A-Type). We will refer to this below as the 'AE' (affect/event) component.

The I-Type event, amounting to about half of the utterances in the ESP corpus, can probably be adequately specified by its textual representation alone, and current speech synthesis technology is already capable of and well-suited for such text-to-speech conversion. We will not touch further on such utterances in this paper.

The A-Type utterances are more text-independent and, to predict how one should be realised, we need to know about the speaker listener relationships (both short-term and long-term), the speaker-state (with respect to (a) emotions, mood, health, and state-of-mind, and (b) current interest and involvement in the dialogue), and thirdly, the intended effect or pragmatic force of the utterance. Note that 'emotion', which is a commonly-used term in the current speech-technology literature, is relegated to a subcategory rather than a dimension in its own right.

In order to synthesise the A-type utterances, we need to know first who is talking to whom, where, and why. An utterance whose primary function is to display affect will be either of a non-lexical type (typically short repeated monosyllables, such as "yeah-yeah-yeah-yeah-yeah", or "uhuh, uhuhuh") or a common phrase, such as "Hi there, how are you?". These 'social grunts' make up as much as half of the utterances in the corpus.

This display of affect as a speech event can be coded in higher-level terms as a combination of the following three features, or 'SOE' constraints: (i) Self, (ii) Other, (iii) Event, as in Eq. (1) which defines an utterance (U) as (probably uniquely) specified by the realisation of a *discourse event* (E) given context-pair *self* (S) and *other* (O).

$$U = E|(S, O) \tag{1}$$

where the feature *Self* can take different values (representing *strong* and *weak* settings with respect to the dimensions *mood* and *interest* respectively) and the feature *Other* can also take different values (representing *strong* and *weak* settings with respect to the dimensions *friend* and *friendly* respectively (see below)), and the feature *event* represents a

Table 1 Dialogue act labels used in the ESP corpus. The first three columns are hierarchically ordered; the labels in fourth column apply only to the 'response' and 'backchannel' categories.

Direction	Category	Dialogue Act	(Response)
	Questions	Question	
		Y/N Question	agree
		Repetition	understand
		Request	convinced
	Opinions	Opinion	accept
		Compliment	interested
		Desire	not convinced
		Will	uncertain
		Thanks	negative
		Apology	repeat
	Negative	Objection	self-convinced
		Complaint	notice
	Advice	Advice	thinking
(offering)		Command	unexpected
(or)		Suggestion	surprise
(seeking)		Offer	doubt
		Inducement	impressed
	Information	Give Information	sympathy
		Reading	compassion
		Introduce Self	other
		Introduce Topic	exclamation
		Closing	listening
	Greetings	Greetings	
		Talking to Self	
		Asking Self	
		Checking Self	
	Other	Notice	
		Laugh	
		Filler	
		Disfluency	
		Mimic	
		Habit	
		Response*	
		Backchannel*	

discourse move or a speech act (in a perhaps wider and more detailed sense than Searle [31] defined. See for example the list in Table 1).

The feature *Self* refers to (a) the state of the speaker and (b) his or her interest in the content of the utterance. For example, a healthy, happy, person is likely to speak more actively than an unhealthy or miserable one. One who is interested in the topic or highly motivated by the discourse is likely to be more expressive than otherwise.

The feature *Other* refers to (a) the relationships between speaker and hearer, and (b) the constraints imposed by the discourse context. A speaker talking with a friend is likely to be more relaxed than when talking with a stranger, but will also probably be more relaxed when talking informally, e.g., in a pub, than when talking formally, e.g., in a lecture hall.

For ease of implementation in a practical speech synthesis system, the Self and Other features can be simplified to a scale of four values each (e.g., plus/minus active & motivated, and plus/minus friend & friendly, respectively). However, the Event feature (e.g., a greeting such as "Nice day, isnt it!" or "Good morning", "Sleep well?", etc.) which

Table 2 Basic utterance types for the 'AE' component. (see Table 1 for a more complete list of discourse events as determined from annotation of the ESP corpus)

	seeking	offering
I-type	interrogative	declarative
A-type	back-channel	exclamative

is used phatically, i.e., not for its lexical meaning but rather for display of speaker-state and speaker-listener relations, can be selected from a wider range of choices according to the contextual constraints described above.

Table 2 represents the Event feature by a simplified two-by-two matrix. Here, as with our annotation of the ESP corpus, each utterance is first categorised in terms of its directionality, i.e., whether it functions for giving or getting, and then in terms of its modality, i.e., whether primarily of I-type or of A-type.

5. Corpora or Databases for Research

The heading of this section poses a question. To date, by far the majority of speech research has been based on databases, specifically produced for the purpose of illustrating one aspect of interest according to sound Newtonian principles. Very little research (except in the closely related fields of text processing) has been based upon analysis or modelling of corpora. For a long while, this was largely accountable as being due to cost limitations, but this is no longer the case. Perhaps as scientists, or engineers, we have become stuck in a rut, and prefer the relative safety of a controlled database to the savage ferocity of a raw corpus?

5.1 Corpus or Database?

A database is a purpose-built collection of structured tokens; a corpus is a collection of naturally-occurring samples from which a database can be constructed. The difference can be explained as one of top-down vs bottom-up design. A database is controlled; specifically designed and constructed to contain useful tokens, being usually relatively small or constrained in size. The contents of a corpus, on the other hand, are by definition not designed, but collected from pre-existing materials exhibiting no inherent order other than that which emerges from the data themselves. A corpus can be focussed on a single feature (e.g., conversational speech, or sports or financial news) but its content should not otherwise be explicitly controlled or contrived. It must be large enough to be representative and to allow trends and patterns to be found from an analysis, but above all it should be natural — a corpus of read-speech, for example, would be considered suspect if the speech it contained were to be read for the purpose of building the corpus itself, rather than for some other explicit purpose.

5.2 Corpora for Speech Synthesis

So-called 'corpus-based' speech synthesis is perhaps there-

fore a contradiction in terms; the source of units from which it is generated should properly be called a database, since it is usually purpose-built (often read from a prepared script or lists of 'balanced' sentences) and produced with the specific aim of illustrating a predetermined set of (usually phonetic or linguistic) features.

However, corpus-based speech synthesis has a long history, dating from before the nineteen eighties. Until then, synthesis by rule was the dominant paradigm, with low-footprint systems predicting not just the prosody but the entire speech waveform characteristics as well. Olive [32]–[34] and Sagisaka [35] proposed using diphones and non-uniform units respectively, for concatenation, to produce more natural-sounding synthetic speech. By utilising segments of actul recorded speech, the important information coded in the non-linear transitions between the phone centres could thus be incorporated directly, rather than being modelled by interpolation in the synthesis.

Unfortuately, the signal-processing required for the prosodic modification and inter-unit smoothing resulted in a degradation of quality such that the resulting synthesis sounded almost as distorted as that produced by rule. Campbell's subsequent introduction of prosody as a unit-selection parameter for concatenative synthesis resulted in much more natural-sounding speech, but at the cost of a dramatic increase in the size of the source database [36]–[40].

5.3 Corpus Size and Quality

Trends in recent years have leaned toward expanding the scale of speech corpora because a larger-scale corpus enables broader phonological and prosodic diversity, which in turn results in improved sound quality. Commercial labs developing concatenative synthesis systems are understandably very sensitive about the precise details of their technologies and reluctant to give out much information about the size of their corpora, but from behind-the-scenes conversations it appears that a source database containing more than 100 hours of speech would not be considered unusual for a commercial-quality system.

However, increasing the corpus size results in increasing costs for development of speech synthesis systems, increases the footprint of the synthesiser, and reduces the flexibility in number of voices that can be synthesised. Research is therefore being carried out in order to quantitatively clarify the relationship between sound quality of the synthesized speech and corpus scale, and to develop methods of designing speech databases to determine the necessary content of a speech corpus when a target domain and a corpus scale are given [41].

The Japan Science & Technology Agency has recently funded a large corpus collection to produce resources for future synthesis research [44]. Being non-commercial, it can focus on capturing the realities of actual spoken conversations rather than limiting the collection to any one domain or speaking style. Whereas the top-down database design of the commercial systems produces a small and balanced source of speech units, there is a danger that the controlled environment and design of such recordings will limit the ability of the system to producing only formal speaking styles and render it incapable of modelling the characteristics of less formal interactive speech.

6. Synthesising Conversational Speech

Table 3 illustrates the flow of processing in traditional speech synthesis. Table 4 contrasts this with the flow required for synthesising conversational speech. In conventional systems, whether text-to-speech or concept-to-speech, the input is usually text-based, and the difference lies in the degree of annotation available to the synthesiser to help predict the syntactic and semantic relationships that render the text intelligible as speech.

Input is parsed to produce an unambiguous phonetic specification and a set of features that determine the prosodic realisation of the utterance. From this cleaned-up text, a numerical specification of the prosody can be obtained, and this is used as a guide to the unit selection. The prosody prediction is limited to information that can be obtained from the text analysis and any markup that may be provided with the input. Currently (see for example the lat-

Table 3 Flow of processing in traditional speech synthesis (e.g., small CHATR, suitable for I-Type utterances). Input text is parsed for its syntactic and semantic content before prosody is predicted and a waveform generated by concatenation of small (typically sub-word) units.

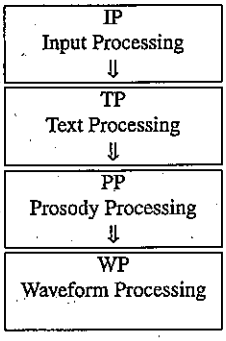
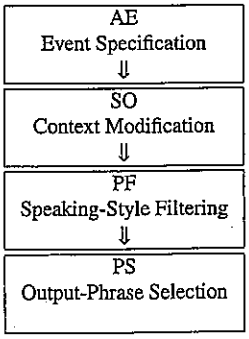


Table 4 Proposed flow for conversational speech synthesis (AESOP — required for A-Type utterances). The combination of Affect and Event is selected as the independent variables in the synthesis process. Context information (Self & Other) modify the result and determine the filters required for selection amongst the multiple phrasal realisations available in the corpus. Output is by replay of the entire phrase-sized corpus utterance.



est w3 proposals for speech markup at [42]) this is limited to low-level features, such as pitch range and speaking rate, and allows little control of voice quality other than can be specified by name, age, and gender (male, female, or neutral). No provision is made for the use of laughter or smiling voice, which, as noted above, are important in interactive situations.

By contrast, the first level of input for synthesising A-Type utterances is the combination of Affect and Event, specifying the AE component. This determines the nature of the utterance, which is preferably not specified directly as text, but rather inferred flexibly by consideration of the 'AE' component in conjunction with the 'SO' layer of context modification. Together, these two levels of processing determine the prosodic and voice-quality filters that constrain the selection of a subset of candidate utterances from the corpus. For example, there may be several thousands of tokens under the category 'greeting' in the corpus, but when constrained by e.g., warm & friendly & interested & relaxed prosodic filtering ('PF') the number of candidates is reduced to several tens of tokens. The final stage of phrase selection concerns continuity; to match the overall characteristics of a given utterance to those of the previous and following phrases so that the output speech does not appear to come from different speakers, as it might if segments from two completely different utterance contexts were selected for contiguous replay.

Since phrase-sized utterances can be extracted whole from the corpus, there is no longer any need to model the linguistic prosodic characteristics, and the 'target-cost' typical of unit-selection is replaced by the selection constraints of the prosody-based PF stage. This results in extremely high naturalness in the 'synthesised' speech, but the mappings between higher-level perceptual (AESO) features and their acoustic characteristics is crucial to effective affect control in the synthesis. The 'target-cost' is not removed, but shifted higher up the perceptual scale from linguistic to paralinguistic levels.

This system has been implemented in Perl and selects utterances with no noticeable delay. Samples of this conversational speech synthesis can be heard (and compared with CHATR synthesis from the same speaker) at http://feast.his.atr.jp/AESOP.

6.1 Specifying Affect & Event

The task of synthesising A-Type conversational speech utterances from a very large expressive speech corpus is very different from that of synthesising an I-Type utterance using a concatenative synthesiser such as CHATR [39]. In the former case utterances can be used intact; whereas in the latter they are made up of small, typically phone-sized, segments selected from other larger utterances. The task of conventional concatenative synthesis (as with the CHATR method) is to select speech segments for concatenation such that they both join well together (i.e., with no perceptible concatenation discontinuities) and at the same time fit the intended prosodic contour(s) closely (i.e., with no perceptible target mismatches).

However, since whole phrase-sized utterances form the basic units of the proposed A-Type conversational speech synthesis, the problem is not one of concatenation discontinuities (since the units are typically separated by a pause or a prosodic break) but rather with determining which of many categorically similar but affectively distinct tokens to use for the actual synthesis; i.e., to minimise the target cost when the target is psychological rather than acoustically determined.

Table 5 summarises the flow of Affect/Event determination. The input-level user-interface component of our AESOP synthesis system is still under active development and is liable to change, but the flow is determined as follows:

First, the precise wording of the Event is preferably undetermined, leaving freedom for the selection of the most appropriate utterance which matches the set of target specifications. These are constrained by the combination of CLASS (greet, confirm, complain, filler, laugh, hedge, accept, decline, etc.,) and VARIANT (happy, sulky, warm, friendly, relaxed, distant, etc.,) in order to select the initial subset of candidate TOKENS from the corpus from which we choose an optimal token according to continuity constraints.

There may not be any randomness at all in human speech variations, all of which can be perceived as meaningful, but an extreme flexibility of expression was found in the ESP corpus. Of approximately 75,000 A-Type utterances, we found an average of 25 tokens each from almost 5,000 classes of 'grunt' (see Table 6), many differing to only a small degree. For example, understanding or realisation might be expressed by 'eh', 'ah', 'oh', or more likely 'eheheh', 'eheheheh', 'hehhehheh', 'hahahahahaha', etc., with repetitions of up to twenty syllables depending on the degree of familiarity, understanding, formality, etc., and by re-

Table 5 Flow of unit-selection control for A-Type conversational utterances: a series of filters produces the final set of candidates.

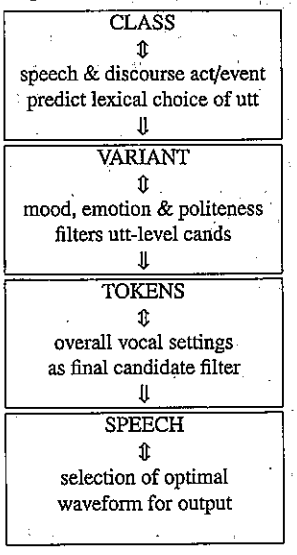


Table 6 Counts of non-verbal utterances in the transcriptions for one speaker in the ESP corpus. Transcribers are encouraged to break utterances into their smallest components (one-per-line) while maintaining sense groups intact. Utterances labelled 'non-lexical' consist mainly of sound sequences and combinations not found in the dictionary, but may include common words such as "yeah", "oh", "uhuh", etc.

number of utterances transcribed	148772
number of unique lexical utterances	75242
number of non-lexical utterances	73480
number of non-lexical utterance types	4492
proportion of non-lexical utterances	49.4%

quiring the user to input a string of text to specify such an utterance (or worse, to be constrained to only only or two variants) would be to ignore a very large part of the corpus. i.e., to fail to faithfully represent the observed variations in the actual situated speech.

It has been confirmed that even without discourse context information, the intended meaning of many of these utterances can be perceived consistently by listeners even when presented in isolation. In many cases, the intentions underlying the utterances can be appropriately and consistently paraphrased even by listeners of completely different cultural and linguistic backgrounds [22], [43].

It is clear from the numbers shown in Table 6 that speakers and listeners must share a protocol for the communication of affective information that can be interpreted in place of, or in line with, the more well-formed I-Type utterances that are produced for the communication of propositional content. That the listener can interpret a grunt (for lack of a better term - since these non-lexical social utterances are not all well described as interjections) in ways that the speaker apparently intended implies that the currently-held linguistic-based assumptions of communication being signalled by semantic elements functioning within a syntactic framework is inadequate for modelling the full range of interactive speech communication. Yet all speech technology and language processing systems are still based largely upon a textual representation of the speech.

6.2 Future Work

Although the entire corpus has been manually transcribed, only a relatively small portion of it (less than 10%) has been annotated by perceptual observation for discourse-act and speaker-state labels. However, we are currently classifying the remaining utterances based on a combination of their text and acoustic characteristics by statistical and semi-automatic methods.

The exact mapping between many of the affective states and the acoustic or textual characteristics of the A-Type utterances is still unknown, and is being discovered by a process of trial-and-error, both through statistical analyses and classification attempts, and by listening to the output of the synthesiser for given input combinations. However, given the flexibility observed in the natural speech, we are optimistic that a similar flexibility can be taken advantage of in the synthesis, since the listener although not a mind-reader is an active participant in the dialogue. Our conversational robot (or speech-impaired person) can serve several brief utterances as in a game of conversational tennis, steering the discourse gradually in the desired direction by use of a series of conversational grunts and common phrases in much the same way as we imprecise humans do in our daily spontaneous interactions.

7. Discussion

This paper opened by contrasting Newtonian approaches to scientific analysis with more recent statistical trends. It claimed that carefully controlled data, with features held constant and only one variable changing at a time, is not representative of living speech, where many features are simultaneously varied to express a highly complex linguistic and interpersonal social message. Large corpora of natural unrestricted speech allow us to examine these multidimensional facets both *in situ* and *in vivo*.

We can learn much about speech by studying its processes in a controlled laboratory environment, but if we are to synthesise speech for an interactive situation, where a synthesiser takes part in a conversation with a human, such as in customer-care applications, aids for the vocally-handicapped, robotics, and games, then we need to incorporate more than just linguistic information and begin to model the interpersonal aspects of spoken interaction. This can only be done if we have access to natural and representative corpora of living speech from ecologically-sound social environments.

The early generations of speech synthesisers employed algorithms to predict the acoustic characteristics of the speech waveforms, and succeeded in mimicking the phonetic properties of speech to the extent that the message was completely intelligible, although not necessarily recognisable as a human voice. Later generations employed recordings of actual speech signals as the source for the output waveforms, concatenating small (typically phone- or diphone-sized) segments and modifying their prosody to match the requirements of the desired output speech. Because of damage caused by the signal processing, the naturalness of the speech was reduced, although its intelligibility was improved.

More recently, speech synthesis systems have made use of very large databases of actual speech, selecting segments for concatenation that embody both the phonetic and the prosodic characteristics required. In this way, the original speaker characteristics are preserved and the speech is not just meaningful but also recognisable as an identifiable human voice. However, most such systems are currently still limited to a single speaking style, as they typically use studio recordings of carefully read speech, often from trained professional speakers, as the source of their waveforms. While adequate for announcements, these source-speech databases include little of the variation encountered in conversational speech, and synthesisers using them are not yet capable of reproducing expressive conversational

speaking styles.

8. Conclusion

This paper has described our approach to the synthesis of conversational speech, based on analysis and use of a very large corpus of natural daily interactions recorded over a period of severa years.

The paper has outlined developments in speech synthesis and shown how it has progressed from the modelling of individual phonemes, through incorporating detailed knowledge of the inter-phonemic transitions, to the use of corpora as a source of large chunks of speech. In parallel with this progression, the role of prosodic information has progressed from that of a boundary and focus marker to one of displaying fine details of speaker state and speaker-listener relationships.

We have shown that as the corpus increases in size and naturalness, so the synthesis process moves from reproduction of sound sequences for the representation of linguistic information, to the reproduction of speaking styles and voice quality for the expression of discoursal and interpersonal affective content.

By using a very large corpus of natural speech as a source for the selection of utterance-sized waveform segments, we have shown that it is possible for a synthesiser to express the same types of information and in the same ways that a human speaker does in normal everyday speech communication. The efficient collection of more such corpora remains as a challenge for future work.

Acknowledgements

This work is partly supported by the Japan Science & Technology Corporation (JST), and partly by the National Institute of Information and Communications Technology (NICT). The author is grateful to the management of ATR for their continuing encouragement and support.

References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," IEEE Signal Process. Mag., vol.18, no.1, pp.32–80, Jan. 2001.

[2] M. Schroder, "Emotional speech synthesis: A review," Proc. Eurospeech, pp.561–564, Aalborg, Denmark, 2001.

[3] P. Ekman, "Basic emotions," in Handbook of Cognition & Emotion, ed. T. Dalgleish and M. Power, pp.301–320, John Wiley, New York, 1999.

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

[5] I.L. Johnson, S. Narayanan, R. Whitney, R. Das, M. Bulut, and C. LaBore, "Limited domain synthesis of expressive military speech for animated characters," Proc. ICSLP 2002, Denver, CO, 2002.

[6] M. Schröder, R. Cowie, E. Douglas-Cowie, M. Westerdijk, and S. Gielen, "Acoustic correlates of emotion dimensions in view of speech synthesis," Proc. Eurospeech 2001, pp.87–90, Denmark, 2001.

[7] A. Iida, N. Campbell, and M. Yasumura, "Design and evaluation of synthesised speech with emotion," J. IPSJ, vol.40, 1998.

[8] I. Lehiste, Suprasegmentals, The MIT Press, Cambridge, 1970.

[9] J. Allen, "Synthesis of speech from unrestricted text," Proc. IEEE, vol.64, no.4, pp.433–442, 1976.

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

[11] I.G. Mattingly, "Experimental methods for speech synthesis by rules," IEEE Trans. AU, vol.16, pp.198–202, 1968.

[12] R. Carlson and B. Granstrom, "A text-to-speech system based entirely on rules," Proc. IEEE-ICASSP76, pp.686–688, 1976.

[13] J. Allen, "Linguistic-based algorithms offer practical text-to-speech systems," Speech Technol., vol.1, no.1, pp.12–16, 1981.

[14] J. Allen, M.S. Hunnicutt, and D.H. Klatt, From Text to Speech: The MITalk System, Cambridge University Press, Cambridge, UK, 1987.

[15] G. Akers and M. Lennig, "Intonation in text-to-speech synthesis: Evaluation of algorithms," J. Acoust. Soc. Amer., vol.77, pp.2157–2165, 1985.

[16] L.C.W. Pols, "Does improved performance also contribute to more phonetic knowledge?," Proc. ESCA Tutorial Day on Speech Synthesis, pp.49–54, Autrans, 1990.

[17] B. Ao, C. Shih, and R. Sproat, "A corpus-based mandarin text-to-speech synthesizer," International Conference on Spoken Language Processing, pp.1771–1774, 1994.

[18] F. Emerard and M. Cartier, "Test d'intelligibilite de systemes de synthese a partir du texte," CNET Report NT/LAA/TSS/427, 1991.

[19] V.J. van Heuven and L.C.W. Pols eds., Analysis and Synthesis of Speech. Strategic Research Towards High-Quality Text-to-Speech Generation, Speech Research, vol.11, Mouton de Gruyter, Berlin, 1993.

[20] T. Dutoit, An Introduction to Text-to-Speech Synthesis, Kluwer Academic Publishers, Dordrecht, 1997.

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

[22] N. Campbell and D. Erickson, "What do people hear? A study of the perception of non-verbal affective information in conversational speech," J. Phonetic Society of Japan, vol.7, no.4, pp.9–28, 2004.

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

[24] N. Campbell, "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.

[25] N. Campbell and P. Mokhtari, "Voice quality: The 4th prosodic dimension," Proc. 15th International Congress of Phonetic Sciences (ICPhS '03), pp.2417–2420, Barcelona, Spain, 2003.

[26] K. Church, "Stress assignment in letter to sound rules for speech synthesis," ACL Proc. 23rd Annual Meeting, pp.246–253, Morristown, NJ, 1985. Association for Computational Linguistics, 1985.

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

[28] N. Campbell, "Extra-semantic protocols: Input requirements for the synthesis of dialogue speech," in Affective Dialogue Systems, ed. E. Andre, L. Dybkjaer, W. Minker, and P. Heisterkamp, pp.221–228, Lecture Notes in Artificial Intelligence Series, Springer, 2004.

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

[30] N. Campbell, "Speech and expression: The value of a longitudinal corpus," Proc. Language Resources and Evaluation Conference (LREC-04), pp.183–186, Lisbon, Portugal, 2004.

[31] J.R. Searle, Speech Acts: An Essay on the Philosophy of Language, Cambridge University Press, Cambridge, 1969.
[32] J.P. Olive, "Rule synthesis of speech from dyadic units," Proc. IEEE-ICASSP77, pp.568–570, 1977.
[33] J.P. Olive, "A scheme for concatenating units for speech synthesis," Proc. IEEE-ICASSP80, pp.568–571, 1980.
[34] J.P. Olive and M. Liberman, "A set of concatenative units for speech synthesis," in ASA*50 Speech Communication Papers, ed. J.J. Wolff and D.H. Klatt, pp.515–518, 1979.
[35] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of nonuniform synthesis units," Proc. IEEE-ICASSP88, pp.679-682, 1988.
[36] W.N. Campbell and C.W. Wightman, "Prosodic coding of syntactic structure in English speech," Proc. ICSLP92, pp.1167–1170, Banff, Canada, 1992.
[37] W.N. Campbell, "Synthesis units for natural English speech," IEICE Technical Report, SP 91-129, 1992.
[38] W.N. Campbell, "CHATR: A high-definition speech re-sequencing system," Proc. Eurospeech '95, Madrid, Spain, 1995.
[39] W.N. Campbell and A.W. Black, "CHATR a multi-lingual speech re-sequencing synthesis system," IEICE Technical Report, SP96-7, 1996.
[40] CHATR Speech Synthesis: http://feast.his.atr.jp/chatr
[41] S. Yamamoto, "Speech translation technologies for real-world applications," ATR Uptodate, vol.3, pp.4–6, 2002.
[42] Speech Synthesis Markup Language Version 1.0 web pages at http://www.w3.org/TR/speech-synthesis/
[43] N. Campbell, "Perception of affect in speech — Towards an automatic processing of paralinguistic information in spoken conversation," Proc. ICSLP 2004, pp.II881–884, 2004.
[44] The Japan Science & Technology Agency, Core Research for Evolutional Science & Technology, 2000–2005.

Nick Campbell is engaged in research as a Project Leader in the Department of Emergent Communication at the ATR Network Informatics Laboratories in Kyoto, and as Research Director for the JST/CREST Expressive Speech Processing project. He received his PhD in Experimental Psychology from the University of Sussex in the U.K. He was invited as a Research Fellow at the IBM UK Scientific Centre, where he developed algorithms for speech synthesis, and was an invited researcher at the AT&T Bell Laboratories. He served as Senior Linguist at the Edinburgh University Centre for Speech Technology Research before joining ATR in 1990. His research interests include large speech databases, concatenative speech synthesis, and prosodic information modelling. Dr. Campbell spends his spare time working with postgraduate students as Visiting Professor at NAIST and Kobe Universities.