Towards Conversational Speech Synthesis;
Lessons Learned from the Expressive Speech Processing Project

Nick Campbell

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

Abstract

This paper discusses some ideas for the requirements and methods of conversational speech synthesis, based on experience gained from the collection and analysis of a very large corpus of conversational speech in a variety of real-life everyday contexts. It shows that because variation in voice quality plays a significant part in the transmission of interpersonal and affect-related social information, this feature should be given priority in future speech synthesis research. Several solutions to this problem are proposed.

Keywords: non-verbal speech, expression of affect, concatenative synthesis, conversational speech corpus, syntax of spoken language

1. Introduction

The JST/ATR Expressive Speech Processing project took place over a period of five years from July 2000 to March 2005 [1]. In that time, 1,500 hours of natural unprompted conversational speech was recorded in a variety of everyday situations using the voices of up to forty paid volunteers as they went about their normal daily activity. Recordings were made directly to DAT or MD using high-quality head-mounted close-talking microphones and all the speech was transcribed manually to form the JST/ATR ESP corpus [2]. A subset of the corpus was further manually labelled to annotate speaking-style and affect-related features.

Given such a large pool of speech samples, including more than 600 hours of speech from one adult female volunteer, a concatenative speech synthesiser was built and tested. The assumption being that given five-years of one person's daily conversations, the system should already contain and so be able to accurately generate most of the speech needed for the sixth or future years from that supply. This turned out not to be the case, but from this work it was discovered that a large amount of the speech was used for expressing interpersonal relationships and affective information [3], rather than propositional content, and considerable effort has since been put into producing a dictionary and a grammar of that mode of nonverbal speech.

2. The Function of Conversational Speech

If speech synthesis is to be developed for conversational applications, such as virtual agents [4], speech translation [5], or 'customer-care' types of two-way spoken interactions, then it will perhaps need to cover the full range of vocal activity encountered in human conversational speech. In other words, it will need to be able to express 'personal feelings' as well as to transmit linguistic information. This will require a degree of prosodic control for which we are currently not well prepared.

Many speech synthesis applications assume a 'broadcast' mode of speech, where the synthesiser speaks and a human listens, with little interaction between the two sides. The focus in broadcast speech is on correctly rendering an input text so that its prosody expresses the syntactic and semantic relations of the component words and their linguistic organisation [6]. Its function is to transmit linguistic information. Contrast this with a conversational mode [7], where the synthesiser also has to take on the role of a listener, providing feedback sounds to signal comprehension (i.e., adequate processing by the dialogue system of the recognised input speech stream), agreement, sympathy, interest, alarm, etc., and their opposites. In this latter 'paralinguistic' mode of speech, the verbal content is limited but its prosodic impact is great. In real interactive speech, laughs and other affect-bursts are common, and 'grunts' take the place of more formal semantics. Phatic communion [8, 9] is as common as (or even more so than) the transfer of linguistic information.

2.1. Non-Verbal Speech Sounds

From the analysis of the ESP corpus, it was learned that approximately half of the conversational-speech utterances are difficult to comprehend from their transcriptions alone. That is, a knowledge of their prosody and voice quality; i.e., of *how* they were spoken, is necessary before an interpretation of their meaning and the speaker's intention can be formed.

A dictionary listing the 100 most frequent utterances in this conversational speech corpus [10] contains words such as "yeah", "okay", "maybe", "gotcha", "uhuh" (i.e., their Japanese equivalents), as well as many laughs, intakes of breath, grunts (such as "ummm", "hmmm", "ooh", etc) and greetings. This list alone is sufficient to cover half of the speech data in terms of utterance frequency. Such non-verbal speech sounds are extremely common. Expanding the list to include similar but less frequent utterances gives at least 2,000 entries excluding laughs, which if we include verbatim (where haha is distinguished from hahaha and hahahaha) involves a further 2,000 types or more.

2.2. Synthesis of Non-Verbal Speech Sounds

Several methods have been tested for the synthesis of such sounds using speech data from the ESP corpus. Being concatenative, they entails little signal processing or prosody or voice-quality manipulation, and simply require the construction of an efficient index to retrieve suitable speech samples from the corpus for replay intact. Although no phone-level concatenation is required for these short complete and self-contained utterances, this method arguably still falls under the umbrella of 'speech synthesis' as it entails the generation of interactive speech utterances by use of a computer system.

The selection of a phatic speech utterance obviously cannot be done just by text alone, as the style and nuance of such speech sounds is much more variable (and informative) than their textual representation. Even something as literal as a greeting, e.g., "Good Morning!", becomes a delicate indicator of speaker-state and speaker-listener relationships through subtle differences in prosody and tone-of-voice, when its phatic role is considered.

Hand-in-hand with the task of selecting appropriately expressive waveform segments is the problem of input; since a computer keyboard (which is limited to the generation of plain text as input) may not be the most appropriate device. Assuming for example that many of our users might prefer to use a portable telephone keypad as their input device of choice, we tested an icon-based menu interface, 'NATR', whereby common conversational utterances could be chosen by toggling the selector button up, down, left, or right (using the thumb) and then pressing a function key to send/synthesise the target speech (see Figure 1). An extended version, 'Chakai', for use with notebook computers with space for occasional free input (shown in Figure 2) has been described elsewhere [11].

Common to both these input devices is a matrix of valency and activation for selection of an appropriate utterance, with icons depicting characteristic features of the utterance in a non-text-based manner. This is because a fundamental assumption of this form of unit selection for conversational speech segments is that the target speech sound is constrained by a set of discourse and interactional features that determine not only its resulting prosody and voice quality, but also the text of the utterance itself. The greeting above is only one form such an utterance might take; when spoken to a close friend it might instead be realised as "Hi!", or clipped down to "mornin' " if the speaker is not feeling too bright. The selection of a phatic utterance should therefore result in a complete and appropriate discourse event, rather than being thought of as determining the prosody and speaking style for a predetermined lexical sequence. This gives the sleection procedure a greater freedom to produce what is most common in the corpus, provided that the labels can effectively constrain selection by representing the factors that generate such an event in the real world.

3. Characteristics of Non-Verbal Speech

In previous work [13] it was proposed that the structure of conversational speech can best be explained as an intermingled sequence of 'wrappers' and 'fillers' such that linguistic content is chunked into small segments that are 'wrapped' by the common and frequently repeated non-verbal speech segments so that both the propositional content and the intended interpretation of the linguistic sequence can be simultaneously conveyed through speech, allowing even a listener unfamiliar with the speech habits of the speaker to be able to interpret the subtle affective changes expressed through micro prosodic and voice-quality variations.

In this paper, we focus more on the lexical, syntactic, acoustic and prosodic characteristics of these 'wrappers' in an attempt to explain how they function and how they might be used to produce natural-sounding utterance sequences for conversational interaction between a person and an agent or agents using speech synthesis.

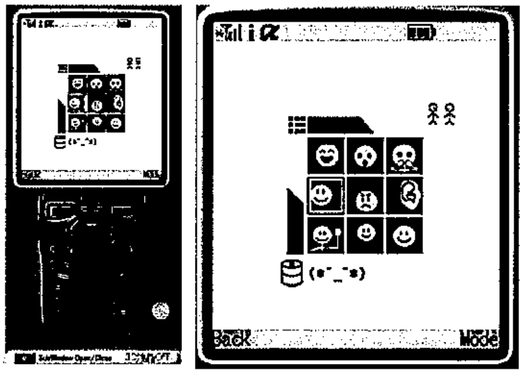

Figure 1: *Input device using portable telephone.*

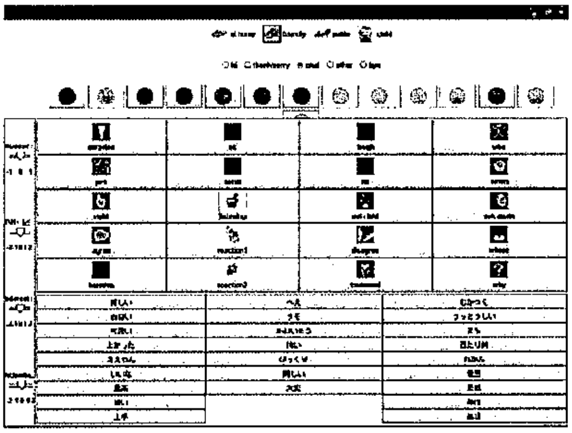

Figure 2: *Input device using notebook computer. Note the space in the centre of the two bottom rows for the input of free text.*

3.1. Wrappers and Fillers - Interaction Devices

Erm, this might seem obvious, but, err, we don't usually use 'wrappers' in text, do we? The previous sentence could better be expressed by seven words (those from 'we' to 'text' inclusive, i.e., what we are calling here the 'fillers'), but nine were added to make it more conversational in style. Contrary to the theory of least effort, it seems that people produce much more speech than 'necessary' (sic) to communicate their intentions. This has been discussed in linguistic science under the competence-performance framework [12], and even today many non-verbal speech sounds are considered to be 'noise'; removed from a recording, not transcribed, covered by a 'garbage-model' in speech recognition, or similarly downgraded and ignored. Errm, one does NOT start a sentence with 'errm'!

And yet these sounds perform a very useful function in discourse. Hesitation is a way of indicating politeness for example, and starting an utterance with 'errm' (or its equivalent) to indicate hesitation is therefore a form of politeness in speech. Stating the obvious, similarly, should not be necessary. How-

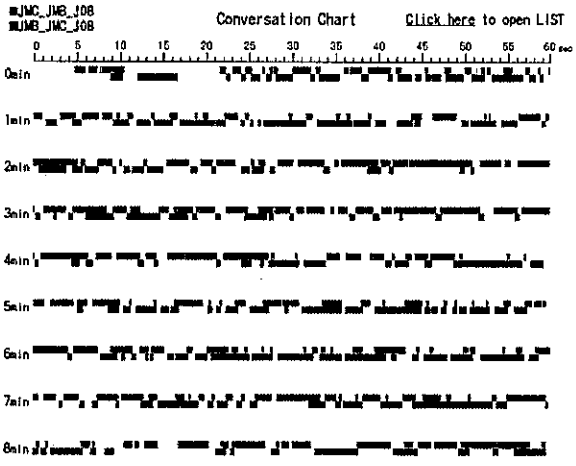

Figure 3: *Speech & silence plots for the first nine minutes of a conversation between two male speakers, JMC and JMB, showing fragmentation of the discourse and progressive but not absolute alternations of speaker dominance. Each line shows one minute of speech, with speaker JMC's speech activity plotted above and that of speaker JMB plotted below. White space indicates lack of speech activity*

ever, "this might seem obvious, but ..." hedges the utterance; it is not redundant but a part of the discourse where the speaker can express affective information, relating to the listener, and to his or her confidence and purpose in speaking. In much the same way "do we?" functions to bring the listener closer into the discourse and to personalise it. It is not a question but a phatic tag.

Furthermore, the very frequency of such tags as "do we?" (here we are including them in the more general term 'wrappers') allows precise variation in expressivity to carry considerable weight of information in the discourse, enabling the speaker to express the degree of belief which the statement is intended to carry. In other words, the linguistic content (i.e., the filling of the utterance) is wrapped in paralinguistic segments that serve to lighten it and to add speaker involvement. This form of speech is limited to conversational styles, and is not found in broadcast modes, where the voice is used solely to portray the content of the text rather than the feelings or attitudes of the speaker.

Such non-verbal (or fringely verbal) use of speech is also particularly common when listening. Active listening demands that the listener chip in frequently to confirm attention, understanding, agreement, etc., and if these phatic sounds are not produced as expected, then most people will simply stop talking. They 'dry-up', asking if the 'listener' is alright perhaps, and the discourse fails as an interactive two-way event.

Figure 3 shows a plot of such two-way activity during a telephone conversation between two people who do not know each other very well. It is probably clear at any given moment in the time sequence who is the dominant speaker, but there is considerable overlap as the listener verbally nods to the speech. Here the same "um" (which is by far the most common utterance in the corpus) can mean yes, no, maybe, just 'I'm listening", 'go on', etc., from differences in intonation, timing, loudness, and voice quality. These are the new challenges for the synthesis of conversational speech. It is not easy to specify the intended variant from text input alone.

Table 1: Results of a principal component analysis of the speech features. We see a decrease in the standard deviation (sd) of the rotated variables as the component number increases, and a decrease in the proportion of the variance (pov) that each component accounts for. By PC7 we note that 82.6% of the cumulative proportion of the acoustic variance (cp) in these data can be accounted for.

```
Importance of components:
      PC1   PC2   PC3   PC4   PC5   PC6   PC7
sd    1.86  1.62  1.35  1.18  1.05  1.00  0.95
pov   0.23  0.17  0.12  0.09  0.07  0.06  0.06
cp    0.23  0.40  0.53  0.62  0.69  0.76  0.82
      PC8   PC9   PC10  PC11  PC12  PC13  PC14
sd    0.87  0.74  0.64  0.59  0.51  0.39  0.31
pov   0.05  0.03  0.02  0.02  0.01  0.01  0.00
cp    0.87  0.91  0.94  0.96  0.98  0.99  1.00

Rotation:
        PC1   PC2   PC3   PC4   PC5   PC6   PC7
fmean -0.35  0.23  0.31 -0.13  0.06 -0.11  0.01
fmax  -0.33  0.15  0.36 -0.11  0.08 -0.14  0.02
fmin  -0.02  0.13 -0.10 -0.52 -0.52 -0.15 -0.11
fpct  -0.20  0.04  0.05 -0.10  0.38 -0.43 -0.57
fvcd   0.19  0.27  0.05  0.55  0.11  0.02 -0.19
pmean  0.03  0.54  0.09  0.11  0.00  0.26  0.01
pmax  -0.24  0.34  0.28 -0.07 -0.11  0.31  0.04
pmin   0.17  0.44 -0.21 -0.12 -0.04  0.09  0.12
ppct   0.05  0.03 -0.09 -0.34  0.67  0.02  0.49
h1h2   0.22 -0.06  0.43  0.15 -0.19 -0.41  0.27
h1a3   0.43 -0.01  0.35 -0.21  0.03  0.01 -0.04
h1     0.42  0.10  0.30 -0.08  0.02 -0.21  0.07
a3    -0.16  0.25 -0.22  0.33 -0.03 -0.46  0.26
dn    -0.11 -0.26  0.37  0.13  0.07  0.37 -0.10
```

3.2. Acoustic features of Wrappers and Fillers

This paper uses the term 'non-verbal' for these speech sounds, but rather than strictly limiting the term to laughs and grunts alone it should be interpreted in its wider meaning to include phrases used more as discourse-gesture than as linguistic content. The example above gave "this might seem obvious" and "do we?" as examples of speech segments that might look like linguistic content but which are actually used more for phatic rather than propositional information transfer. They wrap the linguistic content and give conversational speech its charactertistic 'broken' or so-called 'ill-formed' structure illustrated in Figure 3.

Since these non-verbal wrappers function more to carry prosodic and voice-quality information, it is necessary to categorise them primarily by their acoustic features for unit selection in concatenative conversational speech synthesis. Whereas the prosody of a sentence for broadcast-mode speech synthesis can be largely determined from an analysis of its syntactic, semantic and lexical components and their interactions, the prosody of a phatic grunt for conversational speech synthesis has to be determined independently of (and arguably even before) its lexical composition.

To facilitate the use of acoustic features in unit selection, we used a short program written in Tcl/Tk-Snack [14] to extract the main acoustic and prosodic characteristics of each non-verbal utterance in the corpus to represent its speech waveform as a vector of 14 values (see details in [15]). These include five values (fmean, fmax, fmin, fpct, and fvcd) to represent the pitch contour (fundamental frequency of the speech waveform), four (pmean, pmax, pmin, and ppct) for signal amplitude (power), one for duration, and four to represent spectral characteristics (h1h2, h1a3, h1, a3) of the entire utterance.

The fourteen acoustic and prosodic features thus extracted were then subjected to a principal component analysis to reduce the complexity of the data and to determine the strength of any interactions between the factors. For this, the "prncomp" function in "R" [16] was employed *(pc=prcomp(feats, retx=T, center=T, scale.=T)* which yielded results as shown in Table 1.

3.3. Voice quality and Acoustics

While we see from Table 1 that the principal component analysis allows us to reduce our search space to a smaller number of dimensions, we also note that spectral features rank very highly in explaining the acoustic variation. The data shown in Table 1 were all from the single utterance "umm", the most common word in the corpus, so there is no inherent phonetic variation to be expected that might account for the spectral differences. Instead, the difference in voice quality or breathiness in the speech were used to differentiate between different interpretations of the utterance in the discourse.

"Umm" is used in Japanese, as in English, to mean 'yes', 'I'm listening', 'I understand', 'I agree', 'I don't agree', 'I don't understand', I'm surprised', and so on ..., with each intended meaning unambiguous to the listener but indistinguishable from the text alone. The table shows that more than 50% of this acoustic variance can be accounted for by the first three principal components alone, and that more than 80% can be explained by the first seven. This greatly facilitates search for an appropriate unit.

The table also shows that the first principal component is dominated by h1a3 (i.e., the difference between energy measured at the first harmonic and that measured at the third formant = 0.43), h1 (energy at the first harmonic = 0.42), fmean (mean fundamental frequency of the utterance = 0.35), and fmax (maximum fundamental frequency = 0.32). Power dominates the second principal component, and duration (or speaking rate) the third. Whereas there have been some interesting proposals for modification of spectral tilt in the speech signal (and hence breathiness and 'force' in the speech; see e.g., [17, 18]) the interactions between these four components of prosody is so great that the present author maintains their modification results in unacceptable degradation to the perception of naturalness in the resulting speech and loss of this important voice-quality dimension that is so important for signalling affect and social relationships.

4. KeyTalk

In order to explore the problem of synthesising with a very large number of utterances having a limited number of textual representations but considerable variety in their prosodic expression, and hence in their meaning, we tested a system using a midi-keyboard devise as input for unit selection (see Figure 4).

This system, 'KeyTalk', addresses the problem of grouping related utterances and also of selecting among them by use of a 'force' feature to represent prosodic strength of the utterance. Being coded in the midi language, it allows sequences to be recorded and replayed at a later time or modelled statistically for further synthesis development.

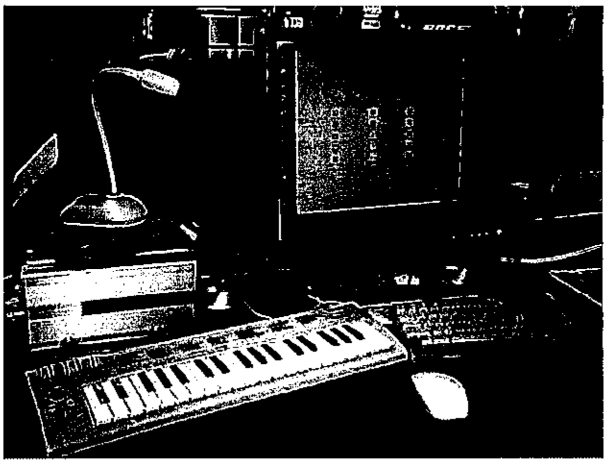

Figure 4: *The KeyTalk setup alongside the NATR conversational speech synthesis interface.*

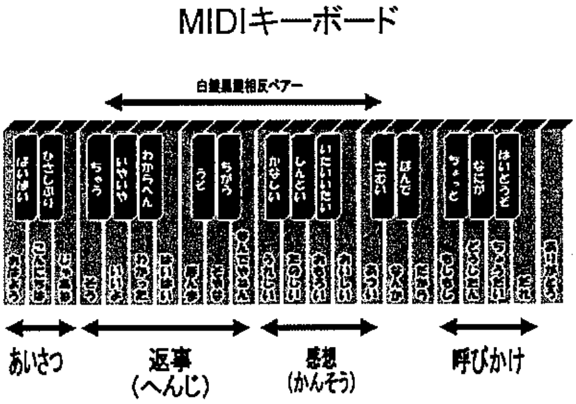

Figure 5: *KeyTalk sample mappings. Groups of keys provide input for related utterances. The keys are touch-sensitive. See text for an exlanation.*

4.1. Grouping Related Utterances

Whereas the data for KeyTalk are complex, the software for the synthesiser itself is very simple. The small piano-like keyboard offers a compact view over the full range of several octaves. Each octave section is grouped into sets of seven and five keys, and alternating within each group are the black and the white keys. The keys are touch-sensitive so a strong keypress will produce a different output-value from a weak keypress, with up to 64 intermediate stages of touch sensitivity.

Each group of keys was mapped deterministically to a group of related and frequently-used conversational utterances, as illustrated in Figure 5 in Japanese. The first group of five keys on the left of the figure represent 'greetings', the next twelve map to 'replies', the next seven map to 'opinions', and those on the right to 'initiating' or 'calling' utterances.

By default, the white keys represent the more positive variants, and black keys their negative equivalents, reflecting the major/minor distinction on a musical keyboard. However, it is not always the case that such simple pairings exist; for example the greetings (white keys) map to morning, afternoon, and evening variants respectively, with the black keys for saying farewell.

Clearly, considerable further work would be required on the selection and grouping of the utterances if this were to be implemented as a commercial system for general use, but as a testbed for experimentation the present working prototype allows 'touch-and-feel' hands-on experience for the selection of individual utterances within a real-time interactive framework.

4.2. Touch-sensitive Selection

The mapping from key to utterance is only a token mapping with no guarantee that the exact word mapped to the key (drawn on the key in Figure 5 for illustration) will be the word that is ultimately spoken by the system. Several modifiers come into action to determine the precise prosody and phrasing of the final utterance. These are governed by global, local, and keypress settings.

Every midi kayboard is also equipped with two roller wheels, one (the 'pitch' wheel) with both positive and negative settings sprung to return to the centre position after each use, and the other, for scalar settings, with no spring, retaining its previous value until further changed. For use as a speech synthesiser, these rollers allow the user to modify the affective profile of each utterance to determine the segment for output. As explained in [11], a three-dimensional control space is posited for conversational speech, whereby the content and style of the utterance are determined from (a) the affective state(s) of the speaker, (b) the character the speaker wishes to display to the listener or conversation partner, and (c) the underlying pragmatic and discoursal intentions of the utterance [19]. This constraint-based unit selection is implemented by ranking candidates for each group of utterances in the database.

As all the speech in the corpus has been transcribed, it is a simple matter to select and group all utterances having an occurrance frequency above a predetermined threshold. These are then ranked according to values of the three principal components described above. The settings of the roller-wheels in combination with the force of the keypress determine which utterance segment from the many ranked candidates will be chosen for playback. The group of candidates from which this selection is made is determined by the key being pressed.

4.3. Evaluation

No formal evaluation has been performed on this system, because each utterance synthesised is a complete and self-contained natural-speech segment. There is no concatenation, except at the phrase level, where utterances are separated naturally by pauses, no prosody modification, and no signal processing. By definition each utterance is natural. Judging how informative it is would be a research issue in itself, because there is as yet no formal grammar of non-verbal utterances against which a sequence of such sounds could be measured. A test of its fun value was carried out in two public demonstrations, one at NAIST (the Nara Institute of Science and Technology) as part of the Open Campus demonstrations in 2006, and again at ATR in the same year as part of the Open House exhibition. In both these cases the keyboard attracted a large number of people and many, especially the younger ones stayed quite a time playing with it, laughing at the sounds that came out, and testing the variety of expression that differences in force of keypress produced.

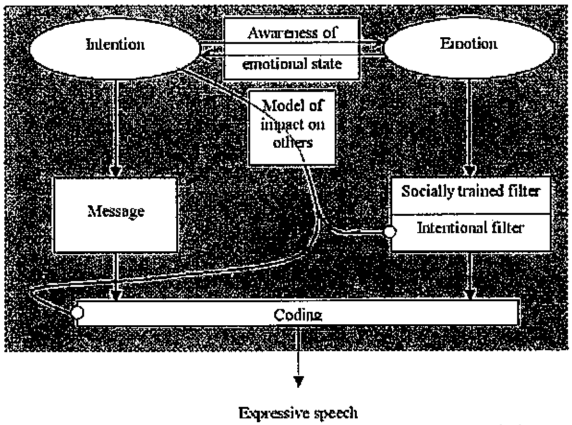

Figure 6: *A model of the constraints (rectangles) and drivers (ovals) underlying the expression of a conversational utterance.*

A different form of evaluation needs to be performed at the level of system design; to find an optimal mapping from keys and clusters of related keys to common utterances in a discourse, and on the mappings between their acoustic characterisatics and the perceived intentions of the speaker, but this requires a formal grammar of spoken language that incorporates non-verbal utterances, and so remains as future work.

5. Discussion

For the system to be of use in an automatic speech synthesiser, a control model must be designed for the generation and integration of non-verbal utterances into the speech stream. One such has been proposed in [21] and is illustrated in Figure 6. Here, two elemental forces are considered as jointly having an influence at the most basic level of the desire to speak. These (marked by ovals in the figure) are hidden and not subject to conscious awareness but must be included in the control model as causative factors.

Below them in the figure are a series of filters (marked by rectangular boxes) representing the constraints that determine the coding of an utterance. This coding is at both the lexical and biomechanical level, resulting in the word sequence and its prosody, simultaneously.

The filters or constraints are of three kinds: (a) the message, i.e the intended pragmatic force of the utterance, what is to be conveyed by the speech, but not yet its precise wording, (b) the social impact of the utterance, on the listener, and in the discourse, and (c) the speaker's character and inhibitions, both trained and innate, as well as the facets of that character to be portrayed (revealed or hidden) in the speech through its content and style. The model assumes emotion and intention to be co-drivers of an utterance, but places most of the control at the level of the constraints.

6. Conclusion

There is growing need internationally for the synthesis of expressive speech, not just in speech translation environments, which are now well developed, but also in the growing area of virtual agents, such as Second Life, where animated beings function in a world of their own, interacting both with each other and with the human sponsors of their communities. The business needs for lifelike conversatioal speech synthesis are great, and very large amounts of real money are already being spent in the virtual communities by a growing number of people across the world.

This paper has described some recent attempts to model the characteristics of conversational speech for use in concatenative speech synthesis, using a very large database of recordings covering a variety of natural environments and interpersonal interactions. Rather than propose a single prototype system, which would be application-specific, it has described several factors that might be taken into consideration in the design of a generic conversational speech synthesis system on the basis of experience gathered from the analysis of the corpus of 1,500 hours of human spoken interactions.

A key theme of the paper is that interactive speech requires different uses of speech prosody from broadcast-mode synthesis, particularly for the expression of affect, interpersonal relationships, and discourse control. Furthermore, in a dialogue system employing conversational speech synthesis, modules will be required for 'active listening' wherein the synthesiser is required to make frequent non-verbal speech sounds to reassure the speaker, to maintain a steady flow of incoming speech, and to control the dialogue. This is an area of discourse which has been little studied, particularly within the engineering and speech technology communities.

The paper has described some of the acoustic features found to be important for the selection of non-verbal speech segments for conversational speech synthesis and has shown that a principal component analysis reduces these to a small manageable number that can be easily ranked and directly used for prosody-based selection of discourse units. The paper has further described some previous attempts at designing novel input devices suitable for use in a conversational environment, both by human users and by computer programs.

7. Acknowledgements

This work is supported by the National Institute of Information and Communications Technology (NiCT), and includes contributions from the Japan Science & Technology Corporation (JST), and the Ministry of Public Management, Home Affairs, Posts and Telecommunications, Japan (SCOPE). The author is especially grateful to the management of Spoken Language Communication Research Labs at ATR for their continuing encouragement and support.

8. References

[1] The Japan Science & Technology Agency *Core Research for Evolutional Science & Technology*, 2000-2005

[2] The JST/CREST Expressive Speech Processing Project homepage can be found at http://feast.atr.jp/

[3] Campbell, N., "Getting to the heart of the matter; speech as expression of affect rather than just text or language", pp 109-118, Language Resources & Evaluation Vol 39, No 1, Springer, 2005.

[4] Second Life: a 3-D virtual world entirely built and owned by its residents. Since opening to the public in 2003, it has grown explosively and at the time of writing is inhabited by a total of 5,788,106 people from around the globe. http://secondlife.com/

[5] Shimizu, T., Ashikari, Y., Sumita, E., Kashioka, H, Nakamura, S., "Development of client-server speech translation system on a multilingual speech communication platform", pp.213-215 in Proc IWSLT, 2006, Kyoto, Japan. 2006.

[6] Crystal, D., "Prosodic Systems and Intonation in English", Cambridge University Press, 1969.

[7] Allwood, J. "An activity based approach to pragmatics". Technical Report (GPTL) 75, Gothenburg Papers in Theoretical Linguistics, University of Goteborg, 1995.

[8] Malinowski, B. "The problem of meaning in primitive languages", pp. 146-152, Supplement to C. Ogden and I. Richards The meaning of meaning. London: Routledge and Kegan Paul. 1923

[9] Jakobson, R., "Linguistics and poetics", pp. 350-77 in Sebeok, T. A.(ed) Style in language. Cambridge, MA: MIT Press, 1960

[10] Campbell, N., "How speech encodes affect and discourse information", pp.103-114 in Esposito, A., Bratani ć, M., Keller, E., and Marinaro, M., *Fundamentals of Verbal and Nonverbal Communication and the Biometric Issue*, IOS Press, Amswterdam, 2007.

[11] Campbell, N., "Conversational Speech Synthesis and the Need for Some Laughter", in IEEE Transactions on Audio, Speech, and Language Processing, Vol 14, No.4, 1171-1179, July 2006.

[12] Chomsky, Noam. 1965. *Aspects of the Theory of Syntax.* Cambridge, Mass.: MIT Press.

[13] Campbell, N., "On the Use of NonVerbal Speech Sounds in Human Communication" in Proc ParaLing'07: Paralinguistic speech - between models and data, Saarbrucken, Germany, 2007.

[14] Sjölander, K., "The Snack Sound Toolkit: a Tcl/Tk library and toolkit for speech signal processing", http://www.speech.kth.se/snack/

[15] Campbell, N., and Nakagawa, A., "'Yes, yes, yes', a word with many meanings; the acoustics associated with intention variation", in Proc ACII '07 (Affective Computing and Intelligent Interaction) Lisbon, Portugal, 2007.

[16] R Development Core Team, "R: A Language and Environment for Statistical Computing", R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, (http://www.R-project.org) 2006.

[17] d'Alessandro, C., & Doval, B. (2003). "Voice quality modification for emotional speech synthesis", pp. 1653-1656. Proc. Eurospeech 2003, Geneva, Switzerland

[18] Kawahara, H., de Cheveigné, A., Banno, H., Takahashi, T. and Irino, T., "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT". Proc. Interspeech2005, Lisboa, pp.537-540, 2005.

[19] Campbell, N., "On the Structure of Spoken Language" in Proc Speech Prosody, Dresden, Germany, 2006.

[20] Mokhtari, P. and Campbell, N., "Quasi-syllabic and quasi-articulatory-gestural units for concatenative speech synthesis", pp.2337-2340 in Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS'03), Barcelona, Spain,. 2003.

[21] Campbell, N., "Expressive / Affective Speech Synthesis", in *Springer Handbook on Speech Processing and Speech Communication*, Benesty, J , Sondhi, M.M., and Huang, Y. (Eds), in Press, Springer, July 2007.