WHAT TYPE OF INPUTS WILL WE NEED FOR EXPRESSIVE SPEECH SYNTHESIS?

Nick Campbell

ATR Human Information Science Laboratories
Kyoto, Japan.

nick@atr.co.jp

Abstract

Speech synthesis is not necessarily synonymous with text-to-speech. This paper describes an implementation for a talking machine that produces multilingual conversational utterances from a combination of speaker, language, speaking-style, and content information, using icon-based input. The paper addresses the problems of specifying the text-content of a conversational utterance from a combination of conceptual icons, in conjunction with language and speaker information. It concludes that in order to specify the speech content (text details and speaking-style) adequately, further selection options for speaker-commitment will be required.

1. Introduction

For unrestricted text-to-speech conversion, the problems of text anomaly resolution and given/new or focus determination are profound. They can require a level of world-knowledge and discourse modeling that is still beyond the capability of most text-to-speech synthesis systems. One implication of this is that the prosody component of the speech synthesiser can only be provided with a default specification of the intentions of the speaker or of the underlying discourse-related meanings of the utterance, resulting in a flat rendering of the text into speech. This is not a problem for the majority of synthesis applications, such as news-reading or information announcement services, but if the synthesiser is to be used in place of a human voice for interactive spoken dialogue, then the speech will be perceived as lacking illocutionary force, or worse, it will give the listener a false impression of the intention of the utterance, leading to potential misunderstandings.

When a synthesiser is to be used in place of a human voice in conversational situations, such as in a communication aid for the vocally impaired, or in call-centre operations, then there is a clear need for the vocal expression of more than just the semantic and syntactic linguistic content of the utterance.

Since the information carried by human speech includes linguistic, para-linguistic, and extra-linguistic details, the listener presumably parses all three sources to gain access to the intended meaning of each utterance. For example, the word 'yes' doesn't always mean yes; when spoken slowly and with a rise-fall-rise intonation, it can instead be interpreted as meaning 'no', or as signalling hesitation, qualifying the interpretation of the lexical content. Similarly, if it is clear from the speech that a speaker is intoxicated (for example) then the listener may be likely to interpret the content of that speech with more caution. Someone speaking with 'an authoritative tone of voice' is more likely to be listened to!

Paralinguistic information, signalled by tone-of-voice, and speaking style, becomes more important as the conversation becomes more personal. Newsreaders and announcers can distance themselves from the content of their utterances by use of an impersonal 'reporting' style of speaking, but customer-care personnel may want to do the opposite in order to calm a client who is complaining, or to reassure one who is uncertain. When speaking with friends, for example, we normally use a different speaking style and tone-of-voice than when addressing a stranger or a wider audience. Speech synthesis must likewise become capable of expressing such differences.

2. Expressive speech

As part of the JST (Japan Science & Technology Agency) CREST (Core Research for Evolutional Science and Technology) ESP (Expressive Speech Processing) Project [1,2], we are collecting 1000 hours of interactive daily-conversational speech, and are building an interface for a CHATR-type synthesizer [3,4] to allow synthesis of speech from the resulting corpus that will be capable of full expressive variation.

Volunteers wear head-mounted close-talking microphones and record their daily spoken interactions to Minidisc devices in blocks of 180 minutes each [5,6]. These samples are then transcribed manually and segmentally aligned automatically from the transcriptions. A large part of the research effort is concerned with the choice of appropriate features for describing the salient points of the interactive speech, and with the development of algorithms and tools for the automatic detection and labelling of equivalent features in the acoustic signal [7,8].

Part of this project includes the development of a communication aid [9,10] and, in particular, an interface for the speedy input of target utterances (the subject of this paper). We are not concerned with text-to-speech processing in this project, and require instead a fully annotated input that is rich enough to specify not just the lexical content of the desired

Figure 1. The Flash web/iPAQ interface

Figure 2. The cell-phone interface (a Java i-Appli).

utterance, but also all aspects of speaking style (including paralinguistic and extralinguistic features) so that speech synthesis can be produced which is appropriate for the discourse context and which will enable the 'speaker' to convey all aspects of the intended meaning.

We are testing our prototypes with disabled users, particularly muscular-dystrophy or ALS patients, who need a speech synthesiser for essential daily communication with friends, family, and care providers [11], but we also envisage business uses of such a system for situations where overt speech may be difficult. For example, a busy executive may want to telephone home to inform her partner that she will be returning later than usual because of a business meeting. She might prefer to use a synthesiser to speak on her behalf, in order not to disturb the meeting. She may also want to convey information regarding the progress of the business deal at the same time. In such a case, the words 'I'll be late tonight' could be spoken with a happy voice to indicate that positive progress is being made. However, if the same message were intended as warning or as an apology, then a happy voice would be quite inappropriate. As humans, we read as much from the tone of voice in such cases as we do from the linguistic message.

The CREST ESP project aims at producing synthesised speech that is able to express paralinguistic as well as linguistic information, and from our analysis of the data collected so far (about 250 hours) we observe that as the interactions become more personal, so the paralinguistic component takes on a greater role in the speech. Utterances become shorter, more common knowledge is assumed, and prosody and voice-quality carry a larger proportion of the information in the message; i.e., the speech becomes more expressive.

3. Icons and utterances

In the case of the business user described above, the use of a keyboard for inputting the text would be highly intrusive into the social situation of a business meeting. Annotating that text for speaking-style information would also be a tedious and time-consuming process. For such situations, we have designed a front-end interface to the synthesiser, for use with a personal assistant or cell phone, so that the speaking style and message can be selected quickly from a menu by toggling buttons. Figure 1 shows a sample screen dump of the GUI interface, programmed in Flash, for use from a web page or personal assistant (iPAQ). Figure 2 shows the equivalent Java-based interface, downloaded to a cell-phone.

3.1. Speech content specification

The buttons on the right of the display in figure 1 are used for selecting speaker, speaking-style, and language respectively. They toggle in a loop and the icon changes to represent the owner of the voice, a face representing the emotion desired for the utterance (currently happy, sad, angry, and normal), and the flag indicating the language (currently only Japanese and English [12]). The equivalent functions on the cell-phone are bound to the numeral keys 1, 2, and 3 on the dial pad. Icons are replaced by text on the phone screen (see figure 2). The icons mapped to the text buttons (left of the display in figure 1) are illustrated in Figure 3 in the form of a table. These are bound to numeral keys 7, 8, and 9 on the cell-phone. 0 is mapped to the 'enter' function to activate the synthesiser. The synthesised speech can be sent directly to the user's device, or redirected to a distant phone.

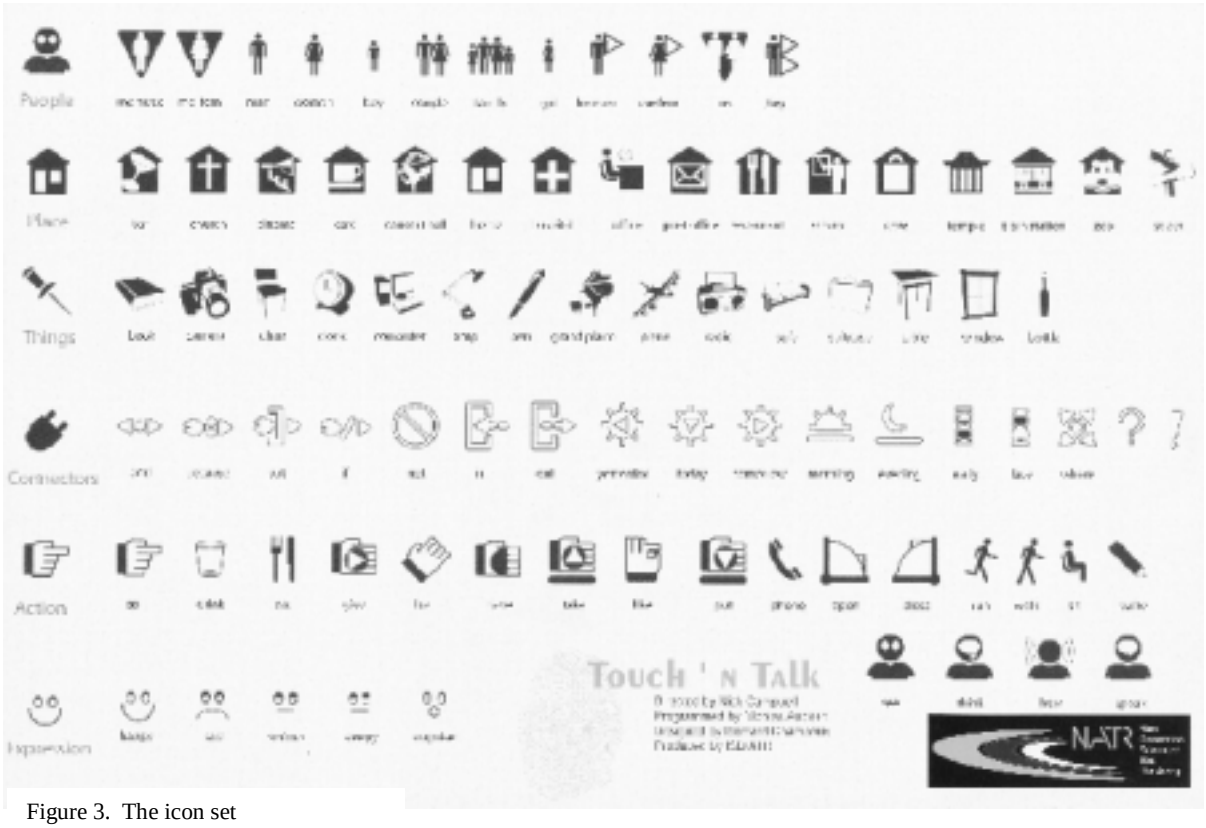

Figure 3. The icon set

The text icons are grouped into five functional classes: 'people', 'places', 'things', 'actions', and 'connectors' (see figure 3). By selecting a combination of these icons, a text to be synthesized can be specified. A basic version of the text appears for confirmation in the display window and can be edited if required. A separate window can be popped up for the entry of additional items in a user-specified word list, e.g., for proper names or slot-fillers. The minimal specification of the utterance allows for modification to the wording of the text according to language and speaking-style settings. Choice of speaker can be programmed to change voice, formality, or personality of the selected speaker, with effects on the wording, prosody, and pronunciation of the utterance.

3.2. Speaking style specification

The texts of all the utterances to be synthesized are produced from components stored in the device (or on the server in the case of cell-phone access) as in domain-specific synthesis. They are finite in number and can be associated with parameter tables specifying e.g., breathiness of the voice, pitch inflections, durational lengthening etc., according to the combination or selection of other parameters by the user.

In the current implementation of this interface, when the user selects an emotion icon, the settings for the speaker-database are changed, and the speech is synthesised using separate source databases, each characterising a different emotion. Work is in progress to merge these individual databases per speaker to enable selection using higher-level descriptors of the speech-style characteristics instead.

The final text generation is hard-coded using a series of conditional and branching operations. All combinations of frequently-used components are exhaustively listed in the source code, and the appropriate prosodic and speaking-style annotations are then added manually. This step is both inelegant and labour-intensive, and we are considering methods of automating the creation of the dictionary component from an analysis of the transcriptions in the ESP corpus.

However, because the text, the translation, the prosody, the voice characteristics, and the speaking style can be all pre-programmed, and do not need to be computed by the synthesiser at run-time, a higher quality of synthesised speech can be guaranteed. The problems of the text-processing and prosody-prediction components have been eliminated from the synthesis process and the brunt of the responsibility rests now on the unit-selection procedures, as a function of source-database coverage and design.

4. Future work

Our experience with the above interface has revealed several aspects of the design that need further consideration. In addition to the database merging and dictionary automation mentioned above, we will also be considering changes to the

'speaking-style' selector. The interface was prepared before we had started analyzing the speech from the conversational corpus, and was designed primarily to facilitate the expression of emotion in synthesised speech.

Analysis of the conversational-speech corpus in terms of 'emotion', using the broad-class labels 'happy', 'sad', 'angry', and 'normal' has proved extremely difficult.

Firstly, the definition of 'normal' appears to be highly context-dependent, as the speaking style varies according to both familiarity with the interlocutor, and type of conversation. Many of the extracts we examined (often just one side of a phone conversation with a friend) were textually very repetitive, but prosodically extremely rich, and varied considerably in their functional meaning. Much of the 'language' consisted of grunts and fillers, being monosyllabic, or repeating the same syllable many times. There is no facility for such back-channeling in the current interface, nor any way of specifying the 'flavour of the grunt' if there were.

Secondly, the 'emotion' labels too seem to be over-simplistic. It is not at all easy to classify a given utterance into one of the above basic classes without first making clear whether we are referring to the speaker's subjective emotional states (both short-term, and long-term) or to the emotional colouring of the utterance itself (and whether intended or not). A dimension of 'control' is needed in addition to the switch for emotion, so that we can distinguish between revealed and intended variants. For example, a schoolteacher might not in fact be angry when speaking in an angry manner to unruly students in the class. Conversely, the person might be feeling extremely angry, but manages for social reasons not to reveal it in the speech. Both of these variants are marked with respect to speaking style.

For the labeling of emotion in the speech database, each utterance must be evaluated separately in terms of such features as the relationships between speaker and hearer (age, sex, familiarity, rank, politeness, etc.), the degree of commitment to the content of the utterance (citing, recalling, revealing, acting, informing, insisting, etc.), the long-term and short-term emotional and attitudinal states of the speaker, the pragmatic force of the speech act, the voice-quality of the utterance (breathy, relaxed, pressed, forced), and so on. The list is not complete. The simplistic notion of a single switch for 'emotion' in a paralinguistic speech synthesiser would appear to need considerable rethinking. The reduction of such complex features to a simple descriptor remains as future work.

5. Conclusion

This paper has presented techniques for the synthesis of conversational speech, expressing paralinguistic information by means of pre-stored annotations on texts. Variants are selected by a combination of icons that represent the basic components of the utterance, the voice, and the speaking style. The implementation is still rudimentary, but experience with the interface is allowing us to design more appropriate ways of specifying the attributes of speech to be synthesised.

For conversational speech synthesis, the specification of the text of an utterance alone is but one small part of the specification of the way in which the utterance is to be produced; and the determination of phrasing and lexical choice will depend on other interacting factors such as speaking style.

The choice of happy, sad, angry, or 'normal' emotions for specifying the speaking style is clearly unsatisfactory, and particular research effort will be needed to determine the most appropriate set of options. We currently believe that at least two further dimensions will be necessary; one specifying the degree of speaker commitment to the content of the utterance ('sincerity'), and one specifying the relationship between the speaker and the hearer ('friendliness').

Acknowledgements

The author is grateful to Bernard Champoux for his help with the graphical components and the iconic language, and to Nicolas Auclerc for his help and advice for the software coding. This work is supported partly by a grant from the Japan Science & Technology Agency under CREST Project #131, and partly by aid from the Telecommunications Advancement Organisation of Japan.

References

[1] JST/CREST Expressive Speech Processing project, introductory web pages at: www.isd.atr.co.jp/esp

[2] Campbell, W.N., "Databases of Emotional Speech", in Proc ISCA (International Speech Communication and Association) ITRW on Speech and Emotion, pp. 34-38, 2000.

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

[4] Campbell, W. N. "Processing a Speech Corpus for CHATR Synthesis". *Proceedings of The International Conference on Speech Processing* 183-186, 1997.

[5] Campbell, W. N., "The Recording of Emotional speech; JST/CREST database research", in Proc LREC 2002.

[6] Campbell, N & Mokhtari, P., DAT vs. Minidisc – "Is MD recording quality good enough for prosodic analysis?", Proc ASJ Spring Meeting 2002, 1-P-27

[7] Campbell, W. N., Marumoto, T., "Automatic labelling of voice-quality in speech databases for synthesis", In Proceedings of 6th ICSLP 2000, pp. 468–471, 2000.

[8] Mokhtari, P, & Campbell, W. N., "Automatic detection of acoustic centres of reliability for tagging paralinguistic information in expressive speech.", in Proc LREC 2002.

[9] Iida, A., Iga, S., Higuchi, F., Campbell, N., Yasumura, M., "A speech synthesis system with emotion for assisting communication", ISCA (International Speech Communication and Assosiation) ITRW on Speech and Emotion, pp.167-172, 2000.

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

[11] Iida, A., Sakurada, Y., Campbell, N., Yasumura, M., "Communication aid for non-vocal people using corpus-based concatenative speech synthesis", Eurospeech 2001.

[12] Campbell,W.N., "Foreign-Language Speech Synthesis", Proceedings ESCA/COCOSDA 3rd Speech Synthesis Workshop, Jenolan Caves, Australia 1998/11/26.