Technology-driven speech synthesis; what does the customer expect?

Nick Campbell
[1] National Institute of Information and Communications Technology
[2] Spoken Language Communication Research Laboratory,
Advanced Telecommunications Research Institute International,
Keihanna Science City, Kyoto 619-0288, Japan.
nick@nict.go.jp/nick@atr.jp

Most of the advances in speech synthesis technology have been driven not by the needs of the market, but by the abilities of the technology. What started out as a reading machine has now developed into a talking machine, but has the original technology really adapted to the needs of its new role? This paper will argue that whereas the technology is already sufficiently competent for the transmission of significant linguistic information, a major component of the prosody of spoken interactions is stilll missing. It provides a model of that component, and shows how small changes in the technology would enable a richer form of speech output, more in line with the needs of conversational interaction and everyday speech.

1. Introduction

At the 3rd International conference of Speech Prosody in Dresden, 2006, there was a noticeable shift in the direction of reported research. In my report to ISCA on SP2006, I noted:

> In addition to the two Special Sessions devoted to Emotion & Affect, there was a plenary oral session on Affective Speech, and a poster session (13 papers) on Prosody & Affect. With a further oral session on Prosody in Pathology & Aging, this seems to indicate a growing interest in the relatively new field of "Social Prosody", extending the scope of prosodic information away from its previous linguistic framework and towards a new interpersonal level of information modelling for spoken communication. [1]

At the 5th Language Resources and Evaluation Conference in Genova, 2006, a similar shift in research emphasis was noted by Calzolari in her introduction to the conference proceedings, summarising the submitted papers:

> [N]ew topics are emerging, linked to subjectivity more than to the 'objective' aspects of meaning, and interestingly this happens both for spoken and written research. I mean topics such as discovery, analysis, representation of sentiments, affect, opinions. This is a new area of research with potentially enormous applicative impact, in areas such as business, marketing, intelligence. The interest for these new topics does not exclude that more 'objective' areas do not present challenges, on the contrary. Despite the progress in the ability to semantically annotate texts, we are far from having 'solved' the problem of 'meaning' or of semantic interpretation of texts. To grasp, manipulate, and effectively use content, both objective and subjective aspects of it, remains the big challenge of our field. Intelligent access to content is thus a goal, maybe a revival — hopefully more successful — of the old Artificial Intelligence with new and more powerful means, i.e. new batteries of tools and resources. [2]

It seems that researchers are becoming more and more aware of the interpersonal aspects of human communication, whether spoken or written, and that they are beginning to process the social and psychological aspects of a message as well as its propositional or linguistic content.

Speech researchers, too, are becoming aware of the multidimensional functioning of speech prosody, and the research focus is now beginning to shift from the function of prosody as a supporter of syntactic and semantic information, signalling phrase-structure and focus, to its broader function of signalling the speaker's standpoint(s) with respect to an utterance, including the affective states, interpersonal relationships, and intended social and linguistic interpretations of the underlying text as part of a discourse. This is a logical progression of interests, because once the narrow linguistic functions of prosody can be well modelled, then the broader social functions will become more apparent; the latter being perhaps derived from the residual of the former in the analysis of a given speech signal.

Speech synthesis, hyowever, typically takes only plain text as input and generates from it a 'suitable' prosodic contour for the utterance (usually limited to duration, and pitch characteristics, and rarely incorporating meaningful amplitude or voice quality variations). This contour is calculated on the basis of syntactic and semantic features derived from the text per se, and has yet to take account of these more subtle prosodic cues needed for social interaction.

The technology has evolved from reading machine to speaking machine as a result of the changing demands of society, but the two tasks are very different. Has the technology also evolved to provide what the customer expects from an interactive speaking device? This paper argues that that is not yet the case, and suggests a direction (and provides a model) by which this may be achieved.

2. Multi-faceted Speech Information

We can describe the extralinguistic and paralinguistic aspects of a spoken message as together providing social, physical, psychological, and interpersonal, information derived from an utterance, and expressed largely through variations in speech prosody.

In order to bettter explain the relation between this 'social prosody' and the more conventional linguistic prosody, a model is here proposed which relates the expression of affect and emotion to the expression of linguistic content in a principled way.

Figure 1 displays the interaction of emotions and intentions in the generation of an expressive speech utterance, as part of a proposed framework for incorporating affect-related information in speech processing. The model arose from extended discussions during and after SP2006, and is in part due to contributions from Sacha Fagel of the TU Berlin. It is still tentative, but is introduced here as a means of illustrating the multi-faceted structure of prosodic information in speech.

2.1. Intentions and Emotions

The model posits two underlying or 'hidden' psychological forces which provide the motivation for a basic communicative event that becomes real in the form of an utterance in a discourse. These are 'Intention' and 'Emotion', drawn within oval shapes in the figure to distinguish them from the more tangible 'Message' and 'Filters' to be discussed below. The lowest box in the figure represents a 'Coding' level of processing which produces commands for the muscles that are used to produce the speech and accompanying facial gestures.

A combination of given 'Intentions' and 'Emotions' represents an underlying socio-psychological state within the speaker which is raw and unbound by social conventions. It can be thought of as an internal force or drive, which is not subject to conscious awareness, and not yet made specific. Here, the term 'emotion' is used in a broad sense to cover the long-term and short-term affective and emotional state(s) of the speaker, including aspects of personality, character, and mood (this is a required disclaimer! [3]). We recognise a semi-conscious process of 'Awareness' relating the hidden and underlying emotions and intentions. Intentions can be triggered by emotions, and emotions can be subdued or amplified intentionally as part of a rational process, such as when a speaker forces herself to smile so that her voice will 'become happier'. However, these processes are above the more tangible level of 'message and filters' which most concerns us here.

2.2. Message and Filters

The message gives form to the underlying intention and constitutes a speech act, a discourse act, and a social event. It may be a greeting, a complaint, provision of information, request for information, etc., and may stand alone or function as a dependent element of a larger discourse. In many cases it will be more phatic than informational in intent.

It is at the level of the message that the utterance begins to take shape, but its linguistic content and prosodic realisation remain indeterminate at this level. For example, a greeting could take the form of "Good Morning", or "Hi!", depending on who is being addressed, on the mood of the speaker, and on the contexts of the discourse (both social and environmental). These details are determined by the settings of the filters.

These filters are socially-trainable. They depend to a large extent on language-specific, culture-specific and subculture-specific aspects. They incorporate such modifiers as politeness constraints and serve to signal attitudinal relationships and interpersonal stance. The filters are shown as bi-level; depending both on social conditioning (above) and intentional control (below). It is at this lower level that the speaker takes into consideration the potential impact of an utterance on the listener (illustrated (not concidentally) in the centre of the figure).

Whereas certain constraints may be ingrained, or determined by society and imprinted in the speaker at an early age, others are more open to conscious selection. For example, while young infants may readily and directly express the emotions they currently feel, older children and adults become more reserved, often concealing their true feelings or masking them for social reasons. A salesperson may wish to portray the proper company image, hiding certain strengths or weaknesses, or a call-centre operator may be required to sound cheerful, even though the displayed 'emotion' may be in conflict with that actually felt by the speaker at the time. This dichotomy provides part of the richness of spoken language and is surely parsed by the listener as part of (or alongside) the message.

Both filter levels function to control (a) what is displayed, and (b) what is concealed in the production of an utterance. They have an effect not just on the selection of lexical items and phrasing, but also on voice quality and posodic aspects of phonation so that the utterance can be parsed appropriately as expresssing the speaker's intentions subject to the prevailing social and psychological states and conditions.

2.3. Coding and Expression

As noted above, selection of lexical items, utterance complexity and length, phrasing, speaking rate and style, etc., can be envisaged as taking place at the lowest level of utterance production, subject to the constraints described above and illustrated in the main part of the figure.

Because there are usually several different ways of phrasing a proposition or eliciting backchannel information, the choice of a particular variant reveals much about the intentions and affective state(s) of the speaker and about the context(s) of the discourse. Both the message and its coding are constrained by the intentions of the speaker, and subject to variations in emotional state and social/intentional constraints on its production.

The figure shows the coding level, which ultimately produces muscular movement sequences, to be fed by two streams of complementary information, as shown by the left and right vertical arrows, both subject to a model of the impact of the utterance on the listener and others. This information can be similarly decoded to reveal not just the linguistic content, but also information about the speaker and the settings of the various filters.

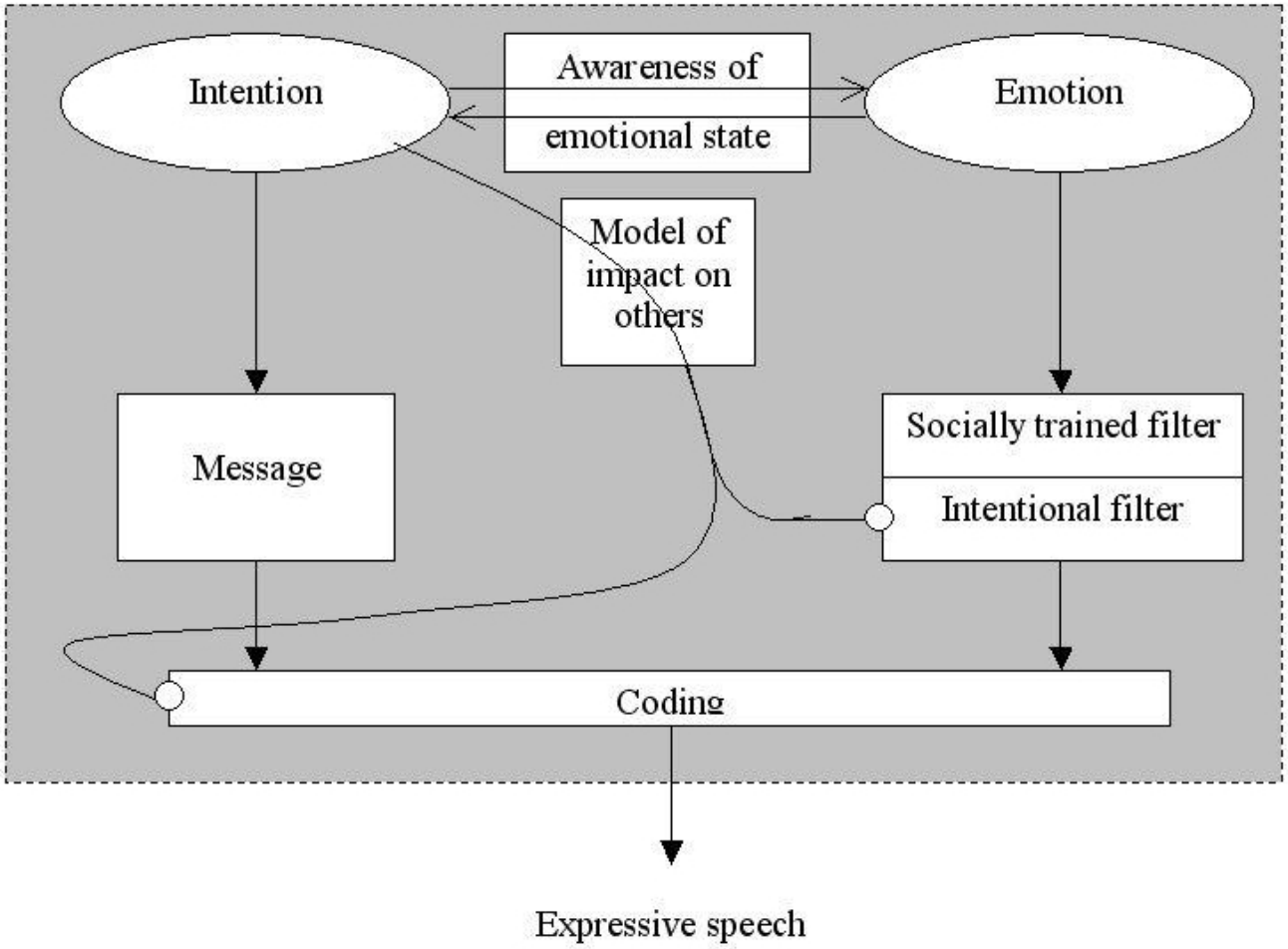

Figure 1: A proposed model to explain the interaction of affective and intentional information in the generation of a speech utterance. The ovals represent 'hidden' processes or states that are not subject to conscious control but which serve as driving forces behind the production of the utterance. These are substantiated in the form of a message and filters, with constraints that are subject to a model of the potential impact on the listener, that determine the muscular coding for the production of the utterance with its resulting prosody and phonetic sequence.

3. Communicative Functions of Speech

The model implies that any given speech utterance contains information related not just to propositional content (if any) but also to speaker-related information, to speaker-hearer relations, and to environmental factors etc., i.e., in addition to the lexical content, or word sequence, an utterance provides both linguistic prosodic and social prosodic information.

These elements are presumably decoded by the listener to reveal the affective and interpersonal information that allows us to understand the speakers position relative to the utterance and thereby to parse its intended meaning from among the many possible candidate interpretations. Whereas the true intentions and emotions of the speaker must remain hidden, much can be inferred about them from the combination of information in the message and in the choice of speaking style (i.e., from the visible effects of the inferred filters), The listener thereby has access not just to the text of the utterance, but also to:

(i) intended meaning(s)
(ii) speaker state(s)
(iii) listener status and relationship(s) to the speaker

This is what is now being covered in the developing studies of social prosody.

3.1. Information Content in Speech

Since I have written in detail elsewhere about the expression of affect in speech [4] it should suffice to summarise briefly here. Humans are primarily social animals; they relate in groups and form close communities and subgroups. Much of human speech is concerned not with the transmission of propositional content or novel information, but with the transfer of affective information; establishing bonds, forming agreements, and reassuring each other of a positive and supporting environment. Or otherwise.

In listening to a spoken utterance, we parse not just its linguistic content, but also the way it has been spoken, voice qualities (including 'tone-of-voice') provide clues to its *intent*, in a way that is complementary to its *content*, to assist in the interpretation of the utterance.

3.2. Prosody in Computer Speech Synthesis

Speech synthesis research is technology-driven; we design according to the perceived needs from the engineer's point of view. There is not yet a strong-enough customer base to allow the technology to evolve in a bottom-up, needs-based, or demand-driven way. There may therefore be a mismatch between what the designers imagine is needed, and what the actual customers want.

From its original goals as a reading machine, the research was focussed on the conversion of written words into spoken sounds, with Grapheme-to-Phoneme conversion, Prosody Prediction, and Waveform Generation as its three main sub-processes. With some notable exceptions (e.g., [5]) there was no representation of emotional content, and the text was considered to contain all of the message.

There is a considerable body of prosody-related research in the field of speech synthesis, but almost all of it (with very few exceptions) is related to the forms of prosody that can be predicted from the text alone. The exceptions largely concern gender-related prosodic differences, or linguistic focus. Existing speech synthesis markup languages (e.g., [6]) allow modification of prosody, but only at the lowest level of mean pitch, phoneme duration, and amplitude. There is very little work yet done on the annotation of text for the expression of the types of affective information described above.

Recently, we have seen an increase of interest in expanding the role of prosody in computer speech synthesis. In particular, there are increasing attempts to include emotion in synthesised speech [7]. Much of the current emotional research is concerned with modelling the prime emotions (anger, fear, joy, sadness, etc.,) [8] but in naturally-occurring spoken interactions (television dramas excepted), the direct expression of raw emotions per se is surprisingly unusual and a more subtle mix of attitudes, as explained by the above model, is more common.

4. What does the Customer Expect?

Computer speech synthesis has mastered linguistic prosody well. Tthere are services widely available for the reading of news and web pages that in some cases cannot be distinguished from natural speech. However, the use of these systems in interactive conversational situations would soon reveal their weaknesses with respect to social prosody. None of them can yet modulate voice quality or speaking style according to differences in the state of the listener, or to reveal suppressed emotions, or even to laugh. Yet in normal human spoken interactions, laughter occurs all the time [9].

If we are to start using speech synthesis in place of the human voice, for e.g., speech-to-speech translation, customer-care services, humanoid robots, or games, etc., then we will have to start modelling the affective information that is so common in human conversational speech. The voice and speaking-style settings reveal much about the discourse contexts and the speakers conscious and unconscious choice of filter settings and constraints. Ordinary people are very used to parsing these types of information in their everyday speech

5. Conclusion

This paper started with a claim that the focus of speech synthesis research to date has been strongly biased by its original reading-machine concept. The paper then presented a model (which is preliminary, and still under development) to show hoqw interactive speech contains information related not just to the message (revealing the speaker's intention) but also to the speaker's affective state(s) and attitudes, and a model of the potential impact of the utterance on the listener.

The implication of the paper is that listeners are aware of both message-related and 'emotion'-related forms of information when listening to interactive (conversational) human speech. The listener to current speech synthesis in a conversational situation, will therefore notice the lack or invariance (inappropriateness?) of the latter forms of information in the speech. This explains the present increase in research related to 'emotion' in speech, but I question whether "emotion" is the best term for this mujlti-faceted stream of personal, interpersonal ans social information that is intertwined with the text of the utterance.q

In summary, the customer of our intaeractive systems probably expects to hear information related not just to the propositional content of the utterance, but also information signalling:

References

[1] SProSIG report, May 2006. ISCA Special Interest Group on Speech Prosody.

[2] Nicoletta Calzolari "Introduction of the Conference Chair", pp I-IV in proc 5th International Conference on Language Resources and Evaluation

[3] Nick Campbell, Laurence Devillers, Ellen Douglas-Cowie, Veronique Auberge, Anton Batliner, and Jianhua Tao, "Resources for the Processing of Affect in Interactions", Panel session, pp. xxiv-xxvii in Proc LREC'06,

[4] Nick Campbell, "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.

[5] Janet Cahn, "The generation of affect in synthesised speech", Journal of the American Voice I/O Society, Vol 8, pp.251-256, 1989.

[6] SSML, The Speech Synthesis Markup Language, www.w3.org/TR/speech-synthesis/

[7] Schroeder, M. "Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions", in Proc. Workshop on Affective Dialogue Systems: Lecture Notes in Computer Science (pp. 209-220. Kloster Irsee, Germany, 2004.

[8] Cowie, R., Douglas-Cowie, E., Cox, C., "Beyond emotion archetypes; Databases for emotion modelling using neural networks", pp 371-388 in Neural Networks 18, 2005.

[9] Nick Campbell "Conversational Speech Synthesis and the Need for Some Laughter", in IEEE Transactionson Audio, Speech, and Language Processing, Vol 14, No.4, July 2006. (in Press)