25. Expressive/Affective Speech Synthesis

N. Campbell

The focus of speech synthesis research has recently shifted from *read speech* towards more *conversational* styles of speech, in order to reproduce those situations where a speech synthesis is used as part of a dialogue. When a speech synthesizer is used to represent the voice of a cognisant agent, whether human or simulated, there is need for more than just the intelligible portrayal of linguistic information; there is also a need for the expression of affect. This chapter reviews some recent advances in the synthesis of expressive speech and shows how the technology can be adapted to include the display of affect in conversational speech.

The chapter discusses how the presence of an interactive and active partner in a conversation can greatly affect the styles of human speech and presents a model of the cognitive processes that result in these differences, which concern not just the acoustic prosody and phonation quality of an utterance, but also its lexical selection and phrasing. It proposes a measure of the ratio of paralinguistic to linguistic content in an utterance as a means of quantifying the expressivity of a speaking style, and closes with a description of a phrase-level concatenative speech synthesis system that is currently in development.

25.1 Overview

Expressive speech synthesis, or the synthesis of affective speech is a relatively new area of research, prompted by a shift in the technology from *reading machines* towards *talking machines* [25.1]. This arose from the need for machines to be able to express more than just the phonetic content of an utterance and to be able to make use of more human-like subtlety in the generation of its prosodic characteristics. Accordingly, speech synthesis research has shifted from *read speech* towards more *conversational* styles of speech. For those situations where speech synthesis is used as part of a dialogue, i. e., where it represents the voice of a cognisant agent, whether human or simulated, then there is a need for more than just the intelligible portrayal of linguistic information; there is also a need for the expression of affect.

Not to be confused with the simpler term ‘emotion’, ‘affect’ as defined in this context refers to the human characteristics that govern the various subtleties of intonation and phrasing, which reveal extralinguistic and paralinguistic information about the speaker, about the speaker’s relationships with the hearer, and about the progress of the discourse and the degrees of mutual understanding attained throughout its progress.

Expressive speech is that which reveals affect. The challenge in synthesizing expressive speech lies in the adequate specification of affective states in the input and in the manipulation or realization of prosodic characteristics to express them in the output speech. This chapter will start by describing the types of speech variation that will be necessary to synthesize, discuss some techniques for eliciting them, and then close by describing some of the approaches that have been taken to tackle the problems of their modeling and realization.

25.2 Characteristics of Affective Speech

Speech per se is a meaningful type of noise which can take place with or without a hearer, but the nature of interactive speech specifically depends on the presence and purposes of a hearer. Broadcast speech represents one extreme of a hypothetical continuum of listener involvement, with romantic murmurings or infant-directed speech perhaps forming the other; currently only the former can be well modeled for synthesis, although there are calls for a more-personal and intimate form of speaking in many current applications of speech synthesis.

Broadcast speech assumes a general audience that is perhaps interested but not necessarily present, and certainly unable to provide any immediate relevant feedback. The focus here is on a unidirectional transmission of information. Conversational speech on the other hand typically requires an active and participating listener. The focus here is on the interaction. The listener may be remote, as in the case of telephone speech, but is required to give constant and timely feedback. The nature of the conversational speech changes according to these feedback reactions from the listener. Classroom speech or lectures can be considered intermediate on this continuum, since the audience is distant but present, able to show comprehension or otherwise, allowing the speaker to moderate and modify the content and speaking style accordingly.

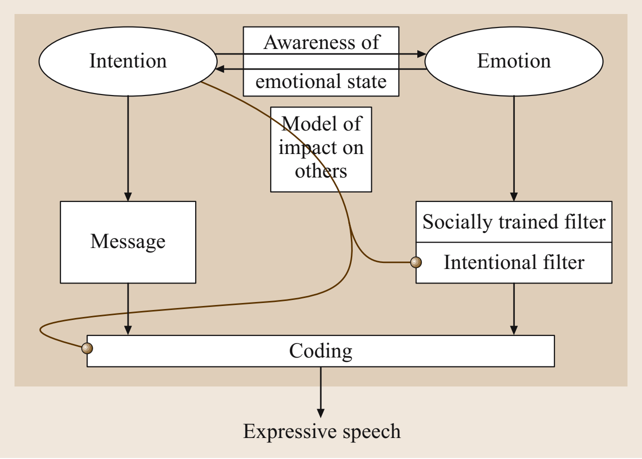


Fig. 25.1 A proposed model to explain the interaction of affective and intentional information in the generation of a speech utterance. The ovals represent hidden processes or states that are not subject to conscious control but which serve as driving forces behind the production of the utterance. These 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

To summarize, affective speech differs from traditional read speech in depending on a model of the listener being incorporated into the speech production process. The task of expressive speech synthesis is first to represent the higher-level interpersonal structures that motivate the speech [25.2], and then to model their lower-level consequences through prosodic control of the output speech signal.

25.2.1 Intentions and Emotions

A representation of the cognitive processes involved in determining a specific speaking style for expressing affect-related information is given in Fig. 25.1, which schematizes the flow of activity that results in expressive speech. The interaction of emotions and intentions in the generation of an expressive speech utterance forms the higher-level part of this framework. In the figure, both *emotions* and *intentions* are represented by ovals, signifying their intangibility; i. e., that they act as a force but are not readily available for identification, control, or conscious appraisal.

The model posits these two underlying or hidden cognitive forces, intention and emotion, as drivers that provide the motivation for a basic communicative event that is to be realized in the form of an utterance in a discourse. They are distinguished from the more-tangible message and filters to be discussed in more detail below. These in turn govern a coding level of processing, which produces lower-level commands for the muscles that produce the speech and accompanying facial or bodily gestures.

A given combination of intentions and emotions represents an underlying sociopsychological state within the speaker, which is raw and unbound by linguistic or social conventions. There may of course be some interaction between emotions and intentions, since intentions can be triggered by emotions, and emotions can in turn 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. These are abstract causes that underly the more-tangible level of message and filters that particularly concerns us here, but they result in subtle differences in force of expression and need to be modeled carefully, since human listeners are particularly sensitive to their variations.

25.2.2 Message and Filters

The message gives form to the underlying intentions and constitutes a speech act, a discourse act, and a social event. It is not yet a text. 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 as much phatic as informational in intent.

It is at the level of the message that the utterance begins to take shape, but its linguistic content and prosodic realization 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 and realized at the level of the coding.

These filters are socially trainable. They depend to a large extent on language-specific, culture-specific and subculture-specific aspects, but can have an intentional element. They incorporate such modifiers as politeness constraints and serve to signal attitudinal relationships and interpersonal stances. The filters are shown as bilevel; depending both on social conditioning (above) and a degree of intentional control (below). It is at this lower level that the speaker takes into consideration the potential impact of an utterance on the listener [as illustrated (not coincidentally) in the center 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 whatever 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 particular strengths or weaknesses, or a call-center 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

1. what is displayed
2. 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 prosodic aspects of phonation so that the utterance can be parsed appropriately as expressing the speaker's intentions subject to the prevailing social and psychological states and conditions. This requires a level of markup on the input far beyond that which is currently specified by the existing W3 speech synthesis markup language (SSML) conventions [25.3].

The sophistication of interactive human speech is a direct result of this multiplicity of cues that taken together contribute the intended interpretation of its linguistic and pragmatic content. The prosody and voice quality of expressive speech also encode subtle information related to the speaker's basic internal emotional states but masked by a more-superficial layer of information related to discourse goals and intentions, and social conventions governing politeness and sincerity of expression. The task of expressive speech processing, whether for synthesis or recognition, is to model the interrelationship of each of these sometimes conflicting sources of information.

25.2.3 Coding and Expression

In this model of expressive speech generation, the determination of lexical items, utterance complexity and length, phrasing, speaking rate and style, etc., is considered as taking place at the lowest level of utterance production, in a constraint-based way, subject to the higher-level 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 states of the speaker and about the contexts of the discourse. Both the text of the message and its prosodic encoding are constrained by the intentions of the speaker, subject to variations in emotional state and social as well as intentional constraints on utterance production.

Figure 25.1 shows the coding level, which ultimately produces sequences of muscular movements, to be fed by two streams of complementary information, as shown by the left and right vertical arrows; both are 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.

25.3 The Communicative Functionality of Speech

The model implies that any given speech utterance contains information related not just to its propositional content, if any, but also to speaker-related and affective information, to speaker–hearer relationships, and to environmental factors etc.; i. e., in addition to the lexical content or word sequence, an utterance provides both linguistic prosody and social prosodic information [25.2, 4]. The former has a long history of being modeled in speech synthesis being predictable largely from the text (see, for example, [25.5–8]), but the latter remains a subject for continuing and intense research, being closely related to database design and corpus collection issues in the case of concatenative synthesis techniques [25.9–11].

These prosodic elements, along with related differences in voice quality, can be decoded by the listener to reveal the affective and interpersonal information that signals the speaker's position relative to the utterance and the discourse, and enables the listener to parse its intended pragmatic meaning from among the many possible linguistic candidate interpretations of the text string. The structure of conversational speech utterances (often called ill-formed) differs considerably from that of their written equivalents not because people are simple and unable to think efficiently in real time, but because speech as a medium has the benefits of multiple layers of prosodic information from which the underlying meanings and intentions of an utterance can be made clear.

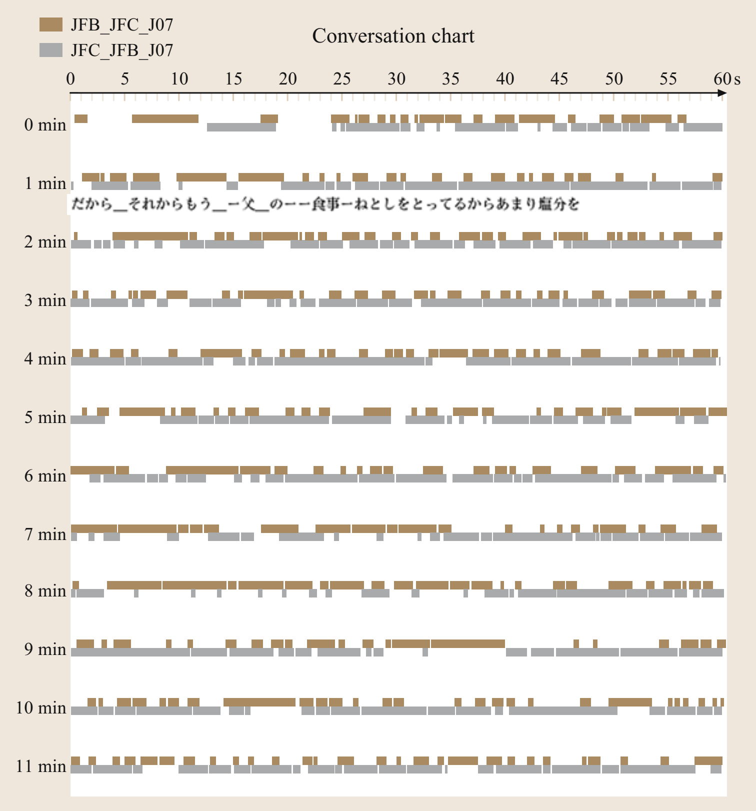


Fig. 25.2 Time-aligned speech activity in a natural telephone conversation, showing how the dominant role is passed from one participant to the other in a not very clear-cut way. The lower speaker (grey) appears to be dominant, but there is considerable overlap as they negotiate their way through the conversation

Figure 25.2 shows the interplay of speaker and listener in a short telephone conversation taken from the expressive speech processing (ESP) corpus [25.12]. It is clear that the roles are often reversed, and that even in listener mode there is considerable speech activity. These backchannel feedback signals provide the speaker with cues to the listener's cognitive state, and comprehension of and agreement with the flow of the conversation as the two mutually develop a joint understanding, or meeting of the minds [25.13].

Conversational speech utterances are not just typically shorter than their written equivalents, they often contain idiomatic phrases, phatic elements, laughs, etc., that illustrate the discourse. Being so frequent and repetitive, these inessential speech noises allow the listener, even one who is unfamiliar with the speaker, to make judgements about the speaker's affective states and discourse intentions.

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:

1. intended meaning(s)
2. speaker state(s)
3. listener status and relationship(s) to the speaker.

This is what is now being covered in the developing studies of social prosody and what is yet to be modeled in expressive speech synthesis.

25.3.1 Multiple Layers of Prosodic Information

Since its original design as a reading machine, speech synthesis research has been focussed on the conversion of written words into speech sounds, using grapheme-to-phoneme conversion, prosody prediction, and waveform generation as its three main subprocesses.

Accordingly, 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., [25.3]) provide for the modification of prosody, but only at the lowest level of mean pitch, phoneme duration, and amplitude. Little work has yet been done on the annotation of text for the expression of the types of affective information described above.

Traditional prosody for speech synthesis is directly related to the text structure of each utterance [25.14, 15], although nowadays it is predicted usually by use of statistical models that have been trained on speech data often unrelated to the voice or personality of the synthesis speaker (e.g., [25.16, 17]). The models provide duration, pitch, and sometimes power values for each segment of the utterance, often through an intermediate phonological representation such as ToBI [25.18, 19].

25.3.2 Text Data versus Speech Synthesis

Spoken language has been extensively studied through the use of corpora for several decades now, and the differences between the types of information that can be conveyed through written texts and those that are signalled through speech are beginning to be well understood. The different types of information that are signalled by different speaking styles are also well understood and are beginning to be modeled in speech technology applications, especially paralinguistic information, which is an essential component of speech communication that is largely carried through modulations of prosody, tone of voice, and speaking style. The more formal the speech, the more constrained the types of paralinguistic information that are conveyed.

At the formal extreme we might consider a public lecture where the speaker is (sometimes literally) talking from a script, to a large number of listeners (or even to a recording device with no listeners physically present) and has minimal feedback from, or two-way interaction with, the audience. This type of spontaneous speech is perhaps the most constrained, and most resembles text. At the most informal extreme, we might consider the mumblings of young lovers. Their conversation is largely phatic, and the words might carry little of linguistic content but are instead rich in feelings. For them, talk is almost a form of physical contact.

There are many steps along the continuum between these two hypothetical extremes of speaking-style variation, and they can be distinguished by the *ratio of paralinguistic to linguistic content*, i. e., the amount of personal information that is included in the speech. The lecture, having almost no personal information and a very high amount of propositional content will result in a very low value of this measure, while the phatic mutterings will score very high.

Speech research depends on the quality and types of the data on which it is modeled. If we are to collect data that contains sufficient examples of natural spoken interactions along the whole range of this continuum of values, then low-scoring material will prove very easy to collect, but most lovers might object strongly to the prospect of having a recording device intruding upon their privacy. Thus, by far the majority of speech corpora that have been used in previous research score very poorly on this scale and as a result the speech that they contain is not very far removed from pure text in its style and content.

25.4 Approaches to Synthesizing Expressive Speech

Expressive speech synthesis is covered under many patents, among them US patent no. 5 559 927 of Sept. 1996, for a *Computer system producing emotionally-expressive speech messages* which describes "a computer system in which the sounds of different speech messages are stored or synthesized, the system being adapted to reproduce a selected speech message and to impart emotional expressivity thereto whose character depends on the user's choice".

But what is emotional expressivity? Take as an example, the website of the Signal Processing Laboratory of the University of the Basque Country (UPV/EHU) in Bilbao [25.20], which announces expressive speech as its title, and which exemplifies what I believe to be a popular and very widespread misconception about expressive speech:

The page states that:

> *Expressive speech synthesis concerns the addition of features related to the emotional state of the (synthetic) speaker as well as the specific characteristics that identify the speaker to the neutral synthetic speech. As such, aspects as emotions (fear, anger, surprise, disgust, happiness, unhappiness ...) and voice quality features (...) should be considered when synthesizing expressive speech.*

It is often taken for granted that expressiveness in speech is the direct result of changes in emotional states, and furthermore that the emotions to be expressed are limited to happiness, sadness, fear, anger, surprise, and disgust (i. e., *Ekman*'s big six [25.21]). This may be an understandable assumption but it is one that does little justice to the sophistication of human interactive communication as described above, unless we are to limit expressive speech synthesis to the realms of storytelling, cartoons, and entertainment [25.22–24].

Ekman was attempting to explain facial expression when he published his seminal work, but the voice is different from the face, and in spoken conversation, what it reveals may not be the emotion so much as the interests of the speaker. In our analysis of a very large corpus of naturally occurring everyday conversations ([25.10,25], see below), these strong emotions accounted for less than 1% of the expressive variability in the speech. However, they account for more than 90% of the related research in speech synthesis.

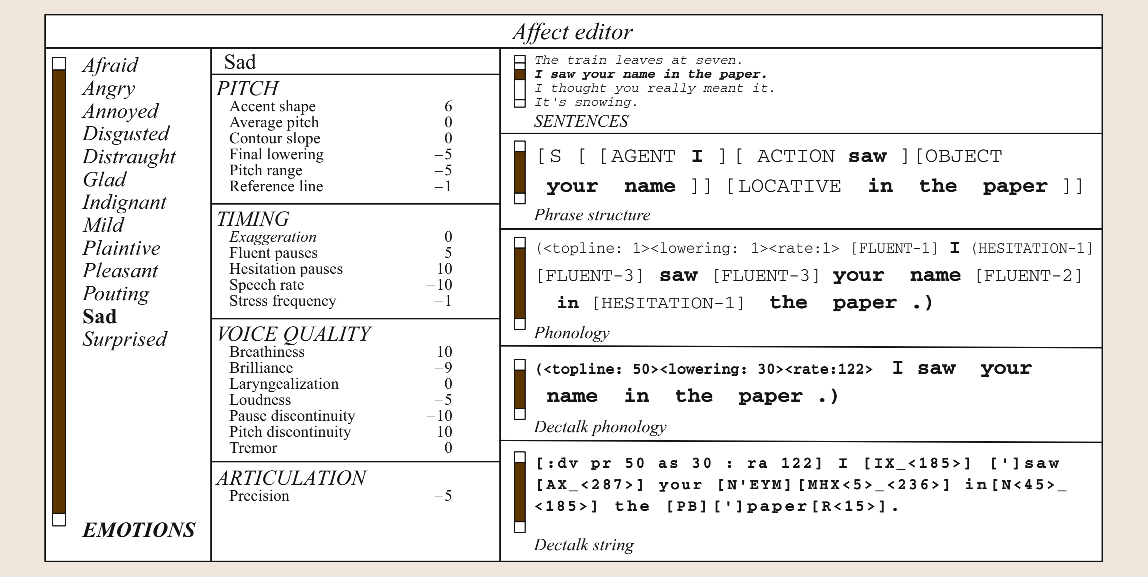
25.4.1 Emotion in Expressive Speech Synthesis

The modeling of expressive speech for speech synthesis perhaps started with the masters thesis of *Janet Cahn* at MIT in 1989 on generating expression in synthesized speech [25.26]. Her work resulted in an affect editor, which modified the prosodic parameters of DecTalk, changing the default settings of pitch, timing, voice quality and articulation according to the emotional states (angry, disgusted, fearful, glad, and sad) that were selected by the user.

Her first paper nearly 20 years ago [25.27] started with the words: "Synthesized speech is readily distinguished from human speech on the basis of inappropriate intonation and insufficient expressiveness", of which perhaps now only the latter part is still true, and she continued: "In synthesized speech intonation makes the message easier to understand; enhanced expressiveness contributes to dramatic effect, making the message easier to listen to," which is still completely true.

It is of particular interest to note that in the 1990 journal paper that described her thesis work [25.28], the simple list of the big six emotions is expanded (Fig. 25.3, which is Fig. 2 in her paper) to include more truly affective states: afraid, angry, annoyed, disgusted, distraught, glad, indignant, mild, plaintive, pleasant, pouting, sad, and surprised, although these were not mentioned in either the thesis or in the 1989 conference paper to the same society [25.29]. This necessary extension from emotion to affect, perhaps encouraged by a sensitive reviewer, has not yet been so readily taken up by the rest of the community of speech synthesis researchers.

A recent paper on expressive speech from AT&T [25.30] for example, describes an experiment



Affect editor

Afraid
Angry
Annoyed
Disgusted
Distraught
Glad
Indignant
Mild
Plaintive
Pleasant
Pouting
Sad
Surprised

EMOTIONS

Sad

PITCH	
Accent shape	6
Average pitch	0
Contour slope	0
Final lowering	−5
Pitch range	−5
Reference line	−1
TIMING	
Exaggeration	0
Fluent pauses	5
Hesitation pauses	10
Speech rate	−10
Stress frequency	−1
VOICE QUALITY	
Breathiness	10
Brilliance	−9
Laryngealization	0
Loudness	−5
Pause discontinuity	−10
Pitch discontinuity	10
Tremor	0
ARTICULATION	
Precision	−5

```
The train leaves at seven.
I saw your name in the paper.
I thought you really meant it.
It's snowing.
```

SENTENCES

```
[S [ [AGENT I ][ ACTION saw ][OBJECT
 your name ]] [LOCATIVE in the paper ]]
```

Phrase structure

```
(<topline: 1><lowering: 1><rate:1> [FLUENT-1] I (HESITATION-1]
[FLUENT-3] saw [FLUENT-3] your name [FLUENT-2]
  in [HESITATION-1] the paper .)
```

Phonology

```
(<topline: 50><lowering: 30><rate:122> I saw your
 name in the paper .)
```

Dectalk phonology

```
[:dv pr 50 as 30 : ra 122] I [IX_<185>] [']saw
[AX_<287>] your [N'EYM][MHX<5>_<236>] in[N<45>_
<185>] the [PB][']paper[R<15>].
```

Dectalk string

Fig. 25.3 The Effect Editor user interface. In this example, the sentence, "I saw your name in the paper." will be spoken with a Sad affect. The parameters of the model and their quantities appear in the middle column, labeled with the curent emotion (Sad). (after [25.28, Fig. 2])

in synthesizing four emotional states – anger, happiness, sadness and 'neutral' [sic] – using a concatenative speech synthesizer. The results were evaluated by conducting listening tests with 33 naive listeners and the synthesized anger was recognized with 86.1% accuracy, sadness with 89.1%, happiness with 44.2%, and neutral emotion with 81.8% accuracy.

Recent developments at MIT for embodied communicative agents are extending Cahn's research, but still primarily from the standpoint of expressing raw emotions. The Kismet project [25.31] explains:

> *emotions have a global impact on speech since they modulate the respiratory system, larynx, vocal tract, muscular system, heart rate, and blood pressure. ... when a subject is in a state of fear, anger, or joy, the sympathetic nervous system is aroused. This induces an increased heart rate, higher blood pressure, changes in depth of respiratory movements, greater sub-glottal pressure, dryness of the mouth, and occasional muscle tremor. The resulting speech is faster, louder, and more precisely enunciated with strong high frequency energy, a higher average pitch, and wider pitch range. In contrast, when a subject is tired, bored, or sad, the parasympathetic nervous system is more active. This causes a decreased heart rate, lower blood pressure, and increased salivation. The resulting speech is typically slower, lower-pitched, more slurred, and with little high frequency energy.*

Since the Kismet goals are to give the appearance of sentience to an inanimate robot, these considerations may be appropriate, but in their recognition of human speech [25.32], they acknowledge concern with the affective states and communicative intentions of the speaker, categorizing incoming speech as one of: soothing, approving, prohibiting, or neutral (i. e., robot-directed). That is, they recognize that some speech is listener-directed rather than speaker-revealing.

Part of the reason for the dominance of raw emotions in expressivity research is the ease of collecting training data. *Campbell* and *Iida* collected an hour each of happy, sad, and angry speech in addition to the standard read-speech to produce an emotional voice for CHATR [25.33, 34]. By switching databases when producing synthesized speech, listeners correctly perceived the intended emotion. However, we soon found that our users, typically speaking-impaired persons such as ALS patients, had less need to express such simple emotions and wanted instead to show interest through tone of voice, or to show concern, warmth, indifference, to be able to communicate their feelings rather than their emotional states.

IBM Research, another team that has an excellent reputation for speech synthesis, has recognized this need for listener-based expression of affect, rather than speaker-based expression of emotion [25.35]. They acknowledge the intentional basis of human communication, and state their text-to-speech research goal to be "to make synthesized speech as intelligible, natural and pleasant to listen to as human speech and have it communicate just as meaningfully". They use concatenative techniques to tune the speech to fit the text content, currently defining expressiveness as distinguishing between: a good news statement, a bad news statement, a yes/no question, and Contrastive emphasis [25.36].

The need to distinguish between urgency and calm in the voice is also realized by the automotive industry. SVOX is continuing to research and develop new ways to extend unit selection algorithms to impart various emotions and expressive voice corpora [25.37]. They claim that:

> *With SVOX SpeechCreate, expressive text-to-speech prompts can be designed and integrated into our customers system. For automotive applications, as an example, an impression of urgency can be included with driving directions that should be followed immediately. In addition, paralinguistic sounds can be added to increase the human-like quality of the speech output.*

There are many similar examples of expressive speech on the web. The interested reader is recommended to visit Felix Burk Burkhardt's site (http://emosamples.syntheticspeech.de) first of all, then to explore the Humaine links [25.38] for a review research related to the expression of emotions in speech.

25.5 Modeling Human Speech

In our own expressive speech research, we first produced a corpus. The Japan Science and Technology Corporation core research for evolutional science and technology (JST/CREST) expressive speech corpus [25.39] was collected over a period of five years by fitting a small number of volunteers with head-mounted high-quality microphones and small minidisc walkman recorders to be worn while going about their ordinary daily social interactions. Further groups of paid volunteers transcribed and annotated the speech data for a variety of characteristics, including speech-act, speaker-state, emotion, relationship to the interlocutor, etc.

Altogether, we collected 1500 h of spontaneous natural conversational speech. All the data were manually transcribed, and about 10% was further annotated. Table 25.1 shows some of the categories that were used for annotation. Speech samples can be listened to at the project website, http://feast.atr.jp/non-verbal/. We found many words and short phrases repeated frequently throughout the corpus; utterances used by the speakers more for their discourse effect than for their linguistic or propositional content. These utterances proved most difficult for the labelers to adequately categorize. They function primarily as backchannel utterances, but also serve to display a wide range of attitudinal and affective states.

We have reported elsewhere [25.4] on studies that measure the extent to which their function can be similarly perceived by different groups of listeners belonging to different cultural backgrounds and languages. In terms of quantity, more than half of the utterances in the corpus were of this predominantly nonverbal type; short words or simple syllables that occurred alone or were repeated several times in succession, often not appearing at all in a dictionary of the formal language, but forming essential components of a two-way spoken interaction.

Our labelers tested several methods of describing these conversational utterances and eventually decided on the three-level system shown in Table 25.1, whereby the following could be distinguished

1. facts about the speaker
2. facts about the utterance, and
3. separate independent evaluations could be made about the information portrayed by differences in the voice quality.

Level 1 describes the state of the speaker, and requires long-term context in order to enable a judgement. It produces an estimation of the discourse purpose of the utterance (see details below), the speaker's emotional states and mood (these labels are free input, those in the table being examples), her interest in the discourse, and finally a label to denote labeler confidence in the current decisions. The numerical labels are forced choice on a scale of high to low (see the lower part of the table) with no default or zero setting.

Level 2 describes the style of the speech, its type and purpose, as can be estimated from a short time window (i. e., with no information regarding discourse context) so that it describes the information available from lis-

Table 25.1 Three levels of labeling for describing each utterance, includingthe use of six-level forced-choice tendency scales

Level 1	**State (about the speaker)**
Purpose	A discourse-act/DA label (see text)
Emotion	Happy/sad/angry/calm
Mood	Worried/tense/frustrated/troubled/ . . .
Interest	A 6-point scale from +3 to −3, omitting 0
Confidence	A 6-point scale from +3 to −3, omitting 0
Level 2	**Style (about the speech)**
Type	Speaking-style label (open class)
Purpose	A discourse-act label (closed class)
Sincerity	Insisting/telling/feeling/recalling/acting/ . . .
Manner	Polite/rude/casual/blunt/sloppy/childish/sexy/ . . .
Mood	Happy/sad/confident/diffident/soft/aggressive/ . . .
Bias	Friendly/warm/jealous/sarcastic/flattering/aloof/ . . .
Level 3	**Voice (about the sound)**
Energy	A six-point scale from +3 to −3, omitting 0
Tension	A six-point scale from +3 to −3, omitting 0
Brightness	A six-point scale from +3 to −3, omitting 0
Level 0	**Labeler**
Confidence	A six-point scale from +3 to −3, omitting 0

Six-point values	**Negative**	**Positive**
Very noticeable	−3	3
Noticeable	−2	2
Only slightly noticeable	−1	1

tening to the isolated speech utterance alone, as distinct from the same utterance situated in a discourse (i. e., we are not interested in whether the speaker is actually, e.g., angry or not, but only in whether the particular segment in question sounds angry). The *sincerity* label describes an important functional aspect of the speech, such as can be distinguished between the verbs 'insisting', 'telling', 'quoting', 'saying', 'feeling', 'recalling', 'acting', 'pretending', etc.

Manner is a bucket category that includes politeness and sexiness (which are not at all mutually contradictory) as well as childishness, sloppiness, etc. to describe the perceived attitude(s) of the speaker towards the listener. This is complemented by mood and bias, of which the former indicates the affective states of the speaker, and the latter his or her attitudes.

Level 3 describes the physical characteristics of the speaker's voice quality and speaking style in perceptual terms.

25.5.1 Discourse-Act Labeling

In order to describe the purpose or function of each utterance, a decision is first made about its directionality, which may be either offering (to the listener) or seeking (from the listener). Utterances are then broadly categorized into seven classes of discourse intentions, including: questions, opinions, objections, advice, information, greetings, and grunts. These category labels are determined by necessity as examples of each appeared in the data. As noted above, the last category accounted for almost half of the utterances in the corpus.

Under the category of 'questions', we use the following labels: WH questions, Y/N questions, repetition requests, and information requests.

Under the category of 'opinions' we use the following labels: opinion, compliment, desire, will, thanks, and apology.

Under the category of 'objections' we use the following labels: objection, complaint.

Under the category of 'advice' we use the following labels: advice, command, suggestion, offer, and inducement

Under the category of 'information' we use the following labels: give information, reading, introduce self, introduce topic, and closing

Under the category of 'greetings' we use the following labels: greeting, talking to self, asking self, checking self.

Under the category of 'grunts' we use the following labels: notice, laugh, filler, disfluency, mimic, habit, response, and backchannel. Response and backchannel utterances are further subcategorized into the following types: agree, understand, convinced, accept, interested, not convinced, uncertain, negative, repeat, self-convinced, notice, thinking, unexpected, surprise, doubt, impressed, sympathy, compassion, exclamation, listening, and other.

25.5.2 Expressive Speech and Emotion

The experience gained from this labeling process has caused us to now rethink some of our original assumptions. We started out by attempting to overcome Labov's observer's paradox on the assumption that long-term exposure to a recording device would eventually cause the wearer to become so familiar with it that it no longer becomes a hindrance to normal spoken interaction, even of a highly personal kind. This has proved to be the case, and is confirmed by the variety of speech that we have collected.

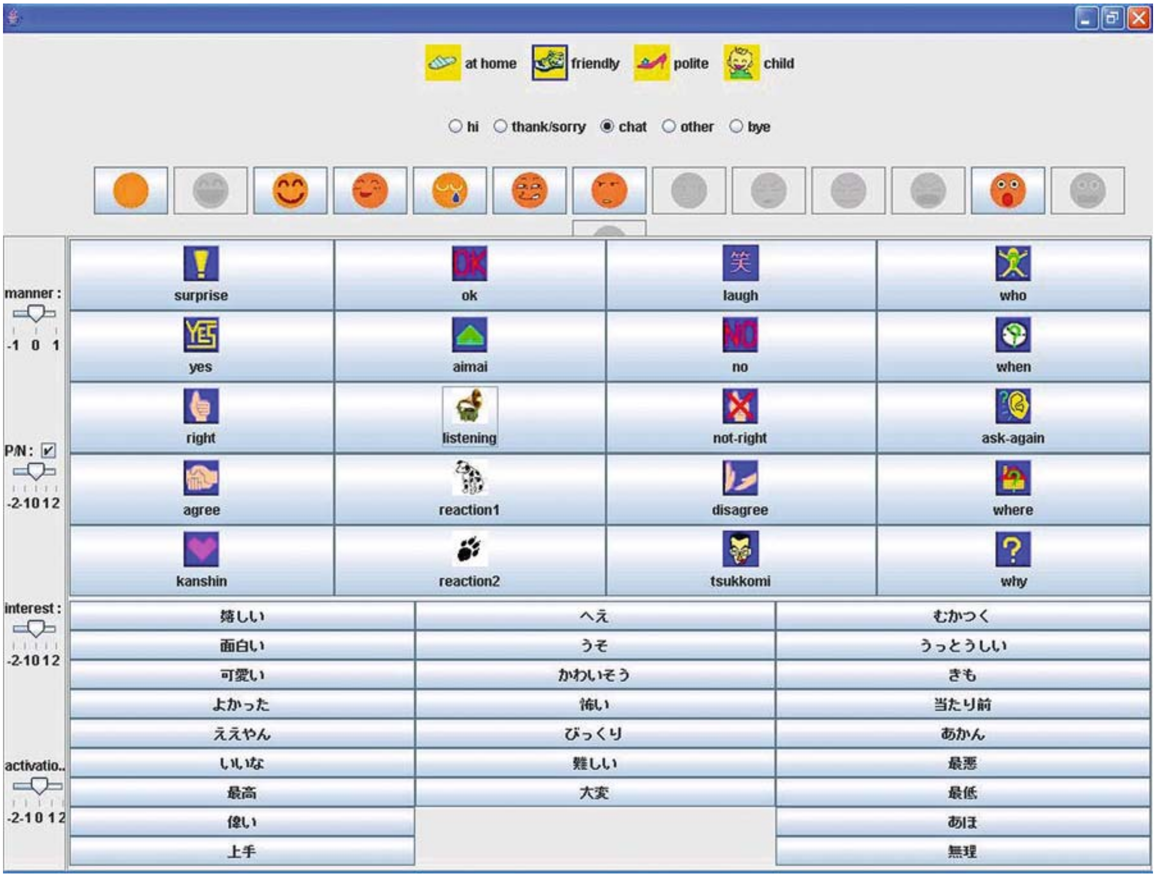


Fig. 25.4 The Chakai conversational speech synthesis interface. By clicking on a speech-act icon, a choice of emoticons is displayed in the *upper section* of the screen according to corpus availability, from which an utterance having the appropriate speech characteristics can be selected. Utterances are selected at random from among those in the same category within the corpus so that subsequent selection of the same combination will provide natural variety without unnecessary repetition

However, another paradox has arisen in its place. We originally believed that we would be able to capture truly natural and spontaneous emotional speech data by having a microphone active and in place before and while the emotional event took place. Instead, we find that by far the majority of our speech material is not marked for emotion as we then conceived it, but that it varies significantly in dimensions better related to affect and attitude, signaling the mood and interest of the speaker, his or her current relations with the listener, and controlling the variable flow of the discourse.

We started out by believing that emotion was the essential component lacking in our speech corpora for technology development, but we now consider that the human dimension that we were looking for is not best described by the term 'emotion' at all. Our data score very highly on the measure of paralinguistic to linguistic content described in the introduction, and are very far from the formal speech of less interactive situations, almost half being nonverbal and affect-related, but they lead us to conclude that the emotional state(s) of the speaker are not always directly expressed, and that social and interpersonal considerations override the supposed link between subjective emotion and displayed affective states. The social aspects of communication therefore take precedence over the blunt expression of feeling, and while the latter can perhaps be determined from an expressive utterance, the multiple levels of information in the former provide a richer source of data to be processed if we are to better understand the person through her speech.

25.5.3 Concatenative Synthesis Using Expressive Speech Samples

With such a large corpus, including more than 600 h from one speaker alone, we were able to test some original approaches to concatenative synthesis of expressive speech. In this case, the units to be concatenated are no longer subsegmental, but can be as large as whole phrases. It might be debatable whether such a concatenative technique is still to be called speech synthesis, but since our unit-selection criteria make use of the labels described above, we would claim that this is still the case.

In parallel with the problem of determining the optimal unit size, is the equivalent problem of how to specify such units for input to the synthesizer. Plain text is no longer appropriate when the intention of the speaker is more important than the lexical sequence of the utterance. Instead, we needed to enable the user to quickly access a given corpus segment by means of higher-level intention-related constraints.

Figure 25.4 shows a recent proposal for Chakai, which is such a speech synthesis interface [25.40, 41] that allows for free input (by typing text into the white box shown at the bottom center) as well as the fast selection of various frequently used phrases and, in addition, an icon-based speech-act selection facility for the most common types of affective *grunt*. The name, not unrelated to CHATR, is composed of two Japanese syllables, meaning tea meeting, an event during which social and undirected chat is common. This format enables linking to a conventional CHATR-type synthesizer for creation of I-type utterances not found in the corpus, while providing a fast, three-click, interface for common A-type utterances which occur most frequently in ordinary conversational speech. Samples (and source code) are available from http://feast.atr.jp/chakai.

The selection of whole phrases from a large conversation speech corpus requires specification not just of the intention of the phrase (greeting, agreement, interest, question, etc.,) but also of the speaker's affective state (as desired to be represented) and the speaker's long- and short-term relationships with the listener at that particular time.

25.6 Conclusion

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 (or other) environment.

In listening to a spoken utterance, human listeners 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.

In speech conversation, both the speaker and the listener are active at all times; the speaker predominantly so, conveying novel information, while the listener conveys feedback information about the flow of the discourse, and about the mutual states of comprehension.

In conversational speech, some of the time is devoted to imparting propositional content, but much of the time is devoted to managing the discourse; eliciting feedback, controlling turn-taking, and expressing affective states and stances.

In the synthesis of expressive speech, we must be able to generate sounds that reproduce all levels of information in the speech signal. At present there are no systems that can do this, but this is an active area of research and one for which there is a strong human need. If we can solve this problem, we will be able to produce a technology that people can relate to, can feel at ease with, and that can process information at levels far deeper than the mere linguistic.

References

25.1 G. Bailly, C. Benoit, T.R. Sawallis (Eds.): *Talking Machines: Theories, Models, and Designs, Reports Papers from the first ISCA Speech Synthesis workshop in Autrans* (North-Holland, Amsterdam 1992)

25.2 N. Campbell: Getting to the heart of the matter; speech as expression of affect rather than just text or language, Language Res. Eval. **39**(1), 109–118 (2005)

25.3 SSML, The W3 Speech Synthesis Markup Language: www.w3.org/TR/speech-synthesis/ (See also the papers from the 2005 SSML Meeting at http://www.w3.org/2005/08/SSML/Papers/)

25.4 N. Campbell, D. Erickson: What do people hear? A study of the perception of non-verbal affective information in conversational speech, J. Phonetic Soc. Jpn. **7**(4), 9–28 (2004)

25.5 I.G. Mattingly: Experimental methods for speech synthesis by rules, IEEE Trans. AU **16**, 198–202 (1968)

25.6 J. Allen: Linguistic-based algorithms offer practical text-to-speech systems, Speech Technol. **1**(1), 12–16 (1981)

25.7 K. Church: Stress assignment in letter to sound rules for speech synthesis. In: *ACL Proc. 23rd Annual Meeting*, ed. by University of Chicago (Association for Computational Linguistics, Chicago 1985) pp. 246–253

25.8 G. Akers, M. Lennig: Intonation in text-to-speech synthesis: Evaluation of algorithms, J. Acoust. Soc. Am. **77**, 2157–2165 (1985)

25.9 N. Campbell: Recording techniques for capturing natural everyday speech. In: *Proc Language Resources and Evaluation Conference LREC-02)* (Las Palmas, Spain 2002) pp. 2029–2032

25.10 N. Campbell: Speech and expression; the value of a longitudinal corpus. In: *Proc. Language Resources and Evaluation Conference* (2004) pp. 183–186

25.11 R. Cowie, E. Douglas-Cowie, C. Cox: Beyond emotion archetypes; Databases for emotion modelling using neural networks, Neural Networks **18**, 371–388 (2005)

25.12 K. Ishimura, N. Campbell: Telephone dialogue data base of JST/CREST expressive speech processing project, Proc. Ann. Conf. JSAI **16**, 147–148 (2002)

25.13 D. McNeill, F. Quek, K.-E. McCullough, S. Duncan, N. Furuyama, R. Bryll, X.-F. Ma, R. Ansari: Catchments, prosody, and discourse, Gesture **1**, 9–33 (2001)

25.14 R. Carlson, B. Granstrom: A text-to-speech system based entirely on rules, Proc. IEEE-ICASSP **76**, 686–688 (1976)

25.15 J. Allen, M.S. Hunnicutt, D.H. Klatt: *From Text to Speech, The MITalk System* (Cambridge Univ. Press, Cambridge 1987)

25.16 K. Hirose, K. Sato, Y. Asano, N. Minematsu: Synthesis of F0 contours using generation process model parameters predicted from unlabeled corpora: Application to emotional speech synthesis, Speech Commun. **46**(3-4), 385–404 (2005-2007)

25.17 A. Sakurai, K. Hirose, N. Minematsu: Data-driven generation of F0 contours using a superpositional model, Speech Commun. **40**(4), 535–549 (2003)

25.18 K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, J. Hirschberg: ToBI: A standard for labelling English prosody, Proc. ICSLP 92 **2**, 867–870 (1992)

25.19 The official ToBI website: http://www.ling.ohio-state.edu/ tobi/

25.20 Webpage of the Signal Processing Laboratory of the University of the Basque Country (UPV/EHU) in Bilbao: http://bips.bi.ehu.es/aholab/TTS/Expressive-Speech-Synthesis.html

25.21 P. Ekman: Basic emotions. In: *Handbook of Cognition and Emotion*, ed. by T. Dalgleish, M. Power (Wiley, New York 1999) pp. 301–320

25.22 K. Sjolander, J. Gustafson: Voice creation for conversational fairy-tale characters. In: *Proc. SSW SYnthesis Workshop* (2005)

25.23 A. Silva, G. Raimundo, C. de Melo, A. Paiva: To tell or not to tell... Building an interactive virtual storyteller. In: *Proc. AISB Symp. Language Speech and Gesture for Expressive Characters* (2004)

25.24 M. Theune, K. Meijs, D. Heylen, R. Ordelman: Generating expressive speech for storytelling applications, IEEE Trans. Audio Speech Language Process. **14**(4), 1137–1144 (2006)

25.25 N. Campbell: Recording techniques for capturing natural everyday speech. In: *Proc. Language Resources and Evaluation Conference* (Las Palmas, Spain 2002) pp. 2029–2032

25.26 J.E. Cahn: *Generating expression in synthesized speech*, M.S. Thesis (Massachusetts Institute of Technology, Cambridge 1989), http://alumni.media.mit.edu/ cahn/emot-speech.html

25.27 J.E. Cahn: From sad to glad: Emotional computer voices. In: *Proc. Speech Tech '88 Voice Input/Output Applications Conference and Exhibition* (New York 1988) pp. 35–37

25.28 J.E. Cahn: The generation of affect in synthesized speech, J. Am. Voice I/O Soc. **8**, 1–19 (1990)

25.29 J.E. Cahn: Generation of affect in synthesized speech. In: *Proc. 1989 Conf. American Voice I/O Society* (Newport Beach, California 1989) pp. 251–256

25.30 M. Bulut, S.S. Narayanan, A.K. Syrdal: Expressive speech synthesis using a concatenative synthesizer, Proc. ICSLP **2002**, 1265–1268 (2002)

25.31 MIT's Kismet (the expressiveness and richness of the robot's vocal modality and how it supports social interaction): http://www.ai.mit.edu/projects/sociable/expressive-speech.html

25.32 MIT Kismet and Affective Intent in Speech: http://www.ai.mit.edu/projects/sociable/affective-intent.html

25.33 A. Iida, N. Campbell, M. Yasumura: Design and evaluation of synthesised speech with emotion, J. Inform. Process. Soc. Jpn. **40** (1998)

25.34 A. Iida, N. Higuchi, N. Campbell, M. Yasumura: Corpus-based speech synthesis system with emotion, Speech Commun. **40**(1–2), 161–187 (2002)

25.35 E. Eide, A. Aaron, R. Bakis, W. Hamza, M.A. Picheny, J.F. Pitrelli: A corpus-based approach to AHEM expressive speech synthesis. In: *Proc. 5th ISCA Speech Synthesis Workshop* (Pittsburgh, USA 2004)

25.36 J.F. Pitrelli, R. Bakis, E.M. Eide, R. Fernandes, W. Hamza, M.A. Picheny: The IBM expressive text-to-speech synthesis system for American English, IEEE Trans. Audio Speech Language Process. **14**(4), 1099–1108 (2006)

25.37 SVOX: http://www.svox.com/Innovation.aspx

25.38 The HUMAINE Portal – Research on Emotions and Human-Machine Interaction: http://emotion-research.net/

25.39 The Expressive Speech Processing project web pages: http://feast.atr.jp/

25.40 N. Campbell: Specifying affect and emotion for expressive speech synthesis. In: *Computational Linguistics and Intelligent Text Processing*, ed. by A. Gelbukh (Springer, Berlin, Heidelberg 2004)

25.41 N. Campbell: Conversational speech synthesis and the need for some laughter, IEEE Trans. Audio Speech Language Process. **14**(4), 1171–1179 (2006)