TOWARDS A GRAMMAR OF SPOKEN LANGUAGE: INCORPORATING PARALINGUISTIC INFORMATION

Nick Campbell

ATR Human Information Science Laboratories
Kyoto 619-022, Japan
nick@atr.co.jp

Abstract

This paper reports on recent developments for the creation and analysis of very large databases of emotional and attitudinally-marked speech for the support of research into concatenative methods for producing synthesised speech which is capable of expressing the range of prosody and phonation styles to emulate human spoken interactions. It addresses the problems of ensuring high spontaneity in the speech corpus while at the same time collecting data that is of high enough audio quality to allow signal analysis by automatic processing techniques. The paper suggests that in order to describe such speech adequately, a new grammar for spoken language will be required.

1. INTRODUCTION

The so-called natural-language grammars have evolved to prescribe the form and structure of written texts. They set out rules or conventions for the combination and ordering of message components so that the recipient of a text may determine the intentions of the writer in order for communication to be effected. Spoken communication, on the other hand, usually takes place under a very different set of constraints. Written texts stand alone, persistent in time, and so must explicitly carry the full information content of their message; whereas speech is often face-to-face, or over a telephone line, with the recipient and the sender of the message simultaneously present. The speech signal is transitory in time, but has the additional dimension of prosodic information to explicitly signal its structure and the component relationships, while at the same time implicitly signaling much about the speaker and about the pragmatic force of the utterance.

There is not yet a formal grammar of spoken language that determines the relationships between all the components of the spoken message in the same way as there is for written text. The two modes of communication share a common linguistic code, but the structuring of the components is greatly different depending on the medium. In addition, the spoken signal is capable of carrying much more implicit information about the state of mind and the intentions of the speaker, communicating extra-linguistic and paralinguistic information as much as linguistic content. In order to prepare for a grammar of spoken language, we first need to collect representative speech data for the analysis. As will be seen below, this is not an easy task.

1.1. Linguistic information

Written text is two-dimensional; the reader can browse it at leisure, scanning up and down the page as well as back and forth along the line, to uncover its structure and content. As speed-reading techniques reveal, it is not constrained to be processed in a linear sequence, even though the words are written in serial order. The structure of the text is revealed through its layout, font styles, headings, underlines, punctuation and paragraphs. Text is designed to be viewed rather than spoken, and the choice of lexis as well as the length and complexity of its sentences is often very different from the spontaneously spoken equivalent. Of course, text can be read aloud, but this media-transform requires considerable mental effort, and only a skilled and practised reader can successfully convey written information by converting it into speech. One reason for the difficulty of reading-aloud is that text is composed to be precise and economical, and to convey maximal information in minimal space. Written text encodes linguistic information according to the strict rules of a grammar.

1.2. Spoken communication

Speech is typically less restricted and less carefully composed. There are rules for formal speaking which can be as restrictive as those for text, but in general conversation, the speech does not follow those rules. The speaker has a larger number of information channels available (even over a telephone, when eye-contact and gesture cannot be used) and adapts the content to the medium, using rhythm, intonation, and voice quality rather than lexical choice to signal the intent of the message.

Because speech is usually interactive, and takes place in real time, it is used to communicate more than just linguistic information alone. Extra-linguistic content such as the sex, age, and condition of the speaker may be obvious from the speech, but these aspects are not usually considered to be part of the message, even though, at one level, the interpretation of the content may differ as a result of them. Paralinguistic details such as the intentions of the speaker, his or her emotions and attitudes, and the pragmatic force of the utterance, are more relevant to the message because they could force a different interpretation of the meaning. To parse such complex multi-tiered and multi-channel information, we need to formulate a framework or grammar to describe its content.

1.3. Paralinguistic information

It is important to distinguish between the content and the function of an utterance. The former can be linguistically defined, and its understanding is independent of any knowledge about the personality of the speaker or the context of the discourse; understanding of the latter requires some knowledge of the situation of the utterance and of the relationships between the speaker and the listener. From a speech technology viewpoint, the psychology of the speaker and the history of the discourse are not easy factors to take into consideration, but we can instead focus on processing the message in the same way that an uninvolved third-party listener might do, using cues from the voice and speaking style alone to interpret from a given pronunciation the intended function of the utterance. We could adopt the standpoint of speech translation and simply ask whether a given utterance would need to be paraphrased or translated differently as a result of the characteristics or manner of its speech production.

In spoken communication, the manner of speaking can carry as much information as the content of the utterance, and the transliteration should vary according to the information carried by the prosody and phonation. Specifically, when processing a spoken utterance, there are cues from the manner of production that must be considered in conjunction with the lexical, semantic, and syntactic variables in order to specify its function or intended meaning. Although a speech recogniser may render the speech into text to produce an accurate representation of the word sequence of an utterance, no recognition system in current use yet takes the prosodic information or the voice quality into consideration to modify the words to represent its intended meaning.

1.4. Prosody and paralinguistics

The simplest example of such functional use of prosody may be seen in the question form of a literal statement, such as: "you're going out tonight?", or "coffee?", the former being a request for confirmation, and the latter an invitation; both would be translated (mistakenly) as declaratives if only the output text of a recogniser was taken into consideration. A more complex example: "coffee!" (with a rise-fall-rise or H*+L,H% intonation), shows the surprise felt by the speaker and carries a more intricate functional message - e.g., "I understand the proposal (or invitation) but I was not expecting it". In both these examples, the intended difference in interpretation can be simply shown by the use of punctuation marks in the text, and is signaled by the use of intonation in the speech.

However, there is another form of expressing speaker attitude or modifying the textual content of an utterance, which has a less obvious correspondence with the punctuation. This involves differences in the manner of phonation of the utterance. For example, "great!" (spoken with a rise-fall or H*+L,L% intonation) is used in colloquial English to show agreement or approval. If spoken with more breathiness in the voice, it can signal more involvement than if spoken with modal phonation, even if the intonation contour is the same. Such deliberate use of phonation style can result in a qualitatively different interpretation of the utterance, and should also be taken into consideration when processing the utterance.

We need to study natural speech data to find out how widespread such functional uses of variation in speaking style can be.

2. DATA COLLECTION AND ANALYSIS

In collaborative work with the Japan Science & Technology Agency, under the auspices of the CREST "Information Processing for Life in an Advanced Media Society", we are collecting data to illustrate the varieties of paralinguistic expression in everyday conversational speech [1].

2.1. Really spontaneous speech

In order to have corpora that are representative of the varieties of speaking styles found in a wide range of everyday situations, the speech should be that of ordinary people naturally expressing various attitudes and emotions in a variety of day-to-day interactive situations.

When a corpus is based on read prompts (e.g., for the study of linguistic aspects of prosody) we can minimise the speakers' personal involvement by focussing their attention on presenting the *form* of the text. The resulting prosody shows only the syntactic and semantic relationships in the text. Questions and statements don't originate from the speaker, but from the text, differentiated by the punctuation alone. The given/new relationships and focus information are similarly inferred – because the speaker is not the *originator* but just an *interpreter* of the text.

In task-based speech collection there is more speaker-involvement, but it is reduced to a paralinguistic minimum. The speaker is not motivated from internal desires, but by the need to perform as requested. Task-based elicitation produces speech with a prosody that signals not just the linguistic framework but also the pragmatic function, since, in a dialogue situation, the listener is as involved as the speaker. A request for information must be signaled as such, in order to obtain a reply without explicit scripting of the speech. Task-based corpora are more natural, but not spontaneous. The speech is unscripted, but the situation is contrived, and the speaker is *cooperating* rather than *operating.*

The need for a balanced scientific design frequently places unnatural requirements on a speech corpus, which render the content less than spontaneous. We can find many examples of such contrived-speech corpora in the literature

2.2. The Observer's Paradox

Corpus design is not the only cause of a lack of spontaneity. In many situations, the presence of an observer can have an influence on that which is being observed. The presence of a microphone (or worse, of a recording engineer) can severely hamper the spontaneity of the speech. The alternative, of surreptitious recording, is ethically questionable (if not illegal) and results in data that cannot easily be shared or published.

In order to overcome this obstacle to natural data collection, we adopted a 'Pirelli-calendar' approach. In 1970 a team of photographers took 1000 rolls of 36-exposure film on location to an island in the Pacific in order to produce a calendar of

twelve (glamour) images. We presume that the reason for this 3000:1 ratio of film to required photographs is that perfect photographs cannot be otherwise guaranteed. By similar 'overkill' reasoning, we assume that if we can record an almost infinite amount of speech, and develop automatic techniques for processing it, to extract only the significant or interesting portions for further analysis, then we will be able to produce a corpus which is both truly representative and of sufficient coverage to allow us to define the full range of prosodic and speaking style variation and to formalize methods to describe its use in human communication.

2.3. Design of a paralinguistic corpus

In order to collect a corpus for the analysis of paralinguistic speech characteristics, we need observer-free recording. The corpus cannot be balanced or designed in the traditional scientific sense because our linguistic concepts may be biased by our views on the *potential* of language use (Chomsky's "competence") and influenced by the text-bound traditions of corpora that are not representative of 'street-speech'.

From a knowledge-base which represents *all* the types of non-verbal information that can be signalled by differences in prosody or phonation of interactive speech (and which categorically alter the perceived effect or 'meaning' of the spoken utterance) we must first produce models of those categories that signal linguistic information, to provide a baseline against which the paralinguistic variations can be contrasted. Speech synthesis front-ends provide a tool for the former, and the difference between predicted and observed speaking characteristics will reveal cues to the latter.

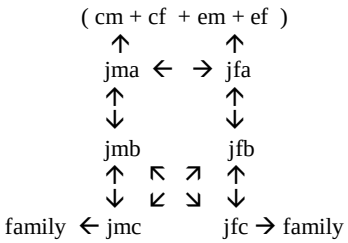
2.4. Data collection

We have to date collected more than 250 hours of unconstrained speech from a range of subjects using two collection paradigms. Both use high-quality head-mounted microphones for recording, but they differ in the recording medium; one using DAT, and the other MiniDisk. The first (recorded on DAT tape) is completely uncontrolled for content, with volunteers telephoning each other at regular intervals to talk freely for half-an-hour per session. Sessions are recorded at weekly intervals for a period of ten weeks. The second (using the lighter and more portable MiniDisk recorders) is an attempt to move beyond laboratory phonology towards 'street phonology', with volunteers recording their daily interactions for extended periods throughout each day.

The composition of the first group is as shown in Table 1. The conversational partners are balanced for familiarity, sex, age, and for ease of communication. All conversations are held in Japanese, but speakers include non-native-language speakers. Arrows show the pairings. Familiarity is minimal at first (except for the family groups) but increases with time throughout the sessions. Only the speech of Japanese native-language speakers will be used for analysis, but all recordings will be preserved.

In a separate data collection, not reported here, we are recording equivalent English and Chinese speech.

Table 1. Recording conditions for the telephone group.



where j = Japanese, m = male, f = female
and e = english native, c=: chinese native
the third letter indicates the pairing, as below:
Group a : cross-cultural difficulties,
Group b : baseline comparison
Group c : talking with family members.

The second group is not so balanced, and only one side of each conversation is recorded. In this group, individuals have agreed to wear ultra-lightweight recording devices while going about their daily work and social interaction. Each MiniDisk allows 160 minutes of continuous high-quality [2] monaural recording of the typically face-to-face interactions. The close-talking head-mounted studio-quality microphone captures the voice of the wearer well, but the voice of the interlocutor is often barely perceptible, so confidentiality of the discourse is assured. Since the interlocutors have not signed release agreements (it would be difficult to make arrangements with such third-parties without intruding on the naturalness of the conversation), only one side of each conversation can be analysed, but because our goal is the analysis of the prosody and phonation style in the conversations, rather than a full conversation-analysis, this is not seen as a problem.

We collect fifty hours of conversation from each subject. The speakers quickly become accustomed to wearing the lightweight recorder, and their speech appears highly natural and typical of normal everyday interactions. All recordings are made in familiar surroundings and with familiar interlocutors. They are of course unscripted and unprompted. The speakers transcribe their own conversations, and have the right to remove any portions of the recordings that they consider to be too personal, but to date, few such deletions have been requested. Instead, respondents have been more concerned that their data must be "much too repetitive" to be of any use to us (!). The text of the utterances is indeed very limited, and reflects the amount of repetition in daily conversational speech, but the prosodic variation is remarkable. The natural repetitions of the lexical, semantic, and syntactic content greatly facilitate both the automatic labelling of the speech, and the comparison of the speaking styles.

We had been concerned that the perceptual-masking-based compression used in the MiniDisk may render the recorded speech unsuitable for signal processing, such as pitch estimation, formant-tracking, and spectral analysis, but comparison tests confirmed that the difference in signal quality between MD and DAT is not significant for these purposes [3] and cannot be heard.

Table 2. Example conversation (family)

00:08:496-00:09:608 C:あったかいなあと思って
00:13:591-00:14:760 B:(? なんちゅうたよ)お父さん
00:15:280-00:15:776 A:え:;上昇調
00:15:824-00:16:648 B:もう来るって
00:16:680-00:17:584 A:もう出ますって
00:17:576-00:17:840 B:はい
00:17:952-00:18:888 C:これ読んだ:;上昇調(疑問調)
00:19:256-00:19:952 B:それ何;疑問調
00:20:200-00:20:583 C:恋
00:21:080-00:24:128 B:*恋はまっまだ*(P 120)これこれ読んだものすごおもしろ*かった
00:21:120-00:21:320 C:*(? 恋)
.

2.5. Data analysis

Table 2 shows an example transcription of a family conversation. While the Japanese may be incomprehensible to the majority of readers of this paper, the table is illustrative of the typical length (and the grammatical complexity, or lack thereof) of each utterance. Sound samples of more example utterances are available at http://feast.his.atr.co.jp [4], and we are confident that even the non-Japanese listeners will be able to hear the intended differences in the interpretation of each utterance from most of the identical-text-pair samples.

There are many repetitions at the lexical level, but each is produced in a different context, and each reveals a different relationship with the listener. For example, the word /hai/ (yes) is pronounced with a variety of meanings, including "yes, I am listening", "yes, I understand", "yes, I agree", as well as "yes, but I don't agree", "what did you say?", "I'm not sure", "I'm not listening to you ", etcetera. Much of the speech is non-verbal; laughs, grunts, and simple one-word utterances are common. These have proved very difficult to transcribe using standard orthography, but they well illustrate how ordinary people speak in real everyday situations, and they have clear communicative intent. They are part of the spoken language.

It is immediately clear from even a cursory analysis of this data that in order to correctly represent the differences in speaker attitude, which are being expressed on each utterance, not just prosodic information but also manner of phonation must be encoded alongside the linguistic information. While a prosodic encoding such as ToBI may be adequate to describe the phonetic structure of the tonal alignments of such utterances, it is not clear that such a system can readily show the attitudinal differences without significant interpretation.

Work is in progress to develop a set of descriptors that capture the necessary relations in order to describe such differences. We have shown that physical measures derived from the speech (such as breathiness of the voice) can map well with pragmatic differences in the utterances [5,6], but more work needs to be done to formalize such relationships. A new "grammar of spoken language" is required, and it will be very different from that which guides the way in which we can form written sentences on a page to be read.

3. DISCUSSION

Speech recognition and speech synthesis technologies currently operate under the assumption that for any given word sequence there need be only one interpretation. The successful speech recogniser puts out the word sequence that represents the text of the input speech, but with no indication of how that sequence is to be interpreted. Similarly, the typical synthesizer will produce only one sound sequence for any given input.

The grammar of spoken language cannot be written without an encoding of the prosody of the speech, and of the pragmatic function of each utterance in an interactive discourse. The data now being collected require novel methods of transcription and novel methods of description, as well as a framework for annotation of the speech, similar perhaps to the markup languages being proposed for speech synthesis, but including functional and attitudinal markers as well as the more mechanical indications of speaking rate and pitch range.

Furthermore, the definition of prosody should be widened to include not just pitch, power, and duration, but also manner of phonation, since this too has a pragmatic effect and is similarly used to modify the intended interpretation of each spoken utterance.

4. CONCLUSION

This paper discussed some of the different types of information that can be signalled through the greater bandwidth of speech, and argued that spoken language needs a qualitatively different formulation of its grammar than that used for written texts. It described the collection of a corpus of really spontaneous speech, and showed that many of the utterances therein may be technically 'ungrammatical' but are nonetheless perfectly intelligible. Most current speech corpora fall into one of two categories: they have high speech signal quality but only illustrate linguistic features, or they illustrate natural speech but have poor speech signal quality. We now have the best of both, but lack a grammar to describe it.

We would appreciate more discussion of the ways of aligning paralinguistic information alongside linguistic content in order to resolve this problem.

5. REFERENCES

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

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

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

[4] Web pages of the *Feature Extraction and Analysis for Speech Technology* project, ATR-HIS.

[5] *Mokhtari, Iida, & Campbell, ICSP 2001, Seoul, Korea.*

[6] Mokhtari & Campbell, in Proc LREC 2002.

* This research was supported in part by a contract with the Telecommunications Advancement Organization of Japan