Chapter 1

EXPRESSIVE SPEECH PROCESSING AND PROSODY ENGINEERING

An illustrated essay on the fragmented nature of real interactive speech

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

Abstract This chapter addresses the issue of human speech communication, focussing not upon the linguistic aspects of speech, but rather on its structure and use in interactive discourse. We show that prosody functions to signal much more than syntactic or semantic relationships.

Keywords: Speech Communication, Affect, Discourse, Synthesis, Recognition.

This paper addresses the issue of expressive speech processing. It attempts to explain a mechanism for expressiveness in speech, and proposes a novel dimension of spoken language processing for speech technology applications, showing that although great progress has already been made, there is still much to be done before we can consider speech processing to be a truly mature technology.

There have been considerable and rapid advances made in the various component technologies over the past ten years, and we now see functioning speech translation devices that are capable of mediating a conversation between people who do not even speak the same language. For fixed-domain applications such as travel or shopping assistance, these devices are capable of recognising speech input in several languages, converting the speech to text, translating the text, and then converting the translated text into speech in a different output language. This successful integration of three separate speech-related technologies, recognition, translation, and synthesis, proves that each has independently reached a degree of maturity in itself, and that all can be used together to model spoken dialogue processes.

However, although the component technologies have been employed successfully within an integrated application, we cannot yet claim them to be fully integrated in a way that models all aspects of spoken interaction. Each has been developed independently of the other, and the implicit assumption behind each component technology is that there is some form of one-to-one mapping between text and speech; i.e., that speech can be rendered as text, text can be manipulated preserving the original content, and that new speech can be generated from existing text. Furthermore there is the underlying assumption that this mapping is sufficient for the processing of spoken language.

In the sections that follow we will show that while the maping may be adequate for the conversion of linguistic or propositional aspects of spoken interaction, it is *not* capable of processing a large part of the social or interpersonal information exchange that takes place in human speech communication, or of recognising and generating the discourse-control signals that speakers use in a conversation. We will examine the role of prosody in spoken language interactions, not from its function as an indicator of syntactic and semantic relationships, but more from the point of view of its role as a social lubricant in mediating human spoken interactions.

Section One will consider the role of prosody in speech communication from a theoretical standpoint, presenting a broader view of prosodic information exchange. Section Two will present some acoustic evidence for the ideas put forward in Section One, and Section Three will suggest some technological applications that might arise from this broader view of spoken language interaction and its related speech processing.

1. Prosodic Information Exchange

The user of a current speech translation system can input a sentence, wait briefly while it is translated, and then hear it reproduced in a foreign language. His or her partner will then be able to reply similarly, producing an utterance in their own language, waiting briefly while it is translated, and then watch the original speaker's reaction while it is synthesised in that person's own language. The processing is in near real-time, so the delays are not long, but the interaction itself is thereby

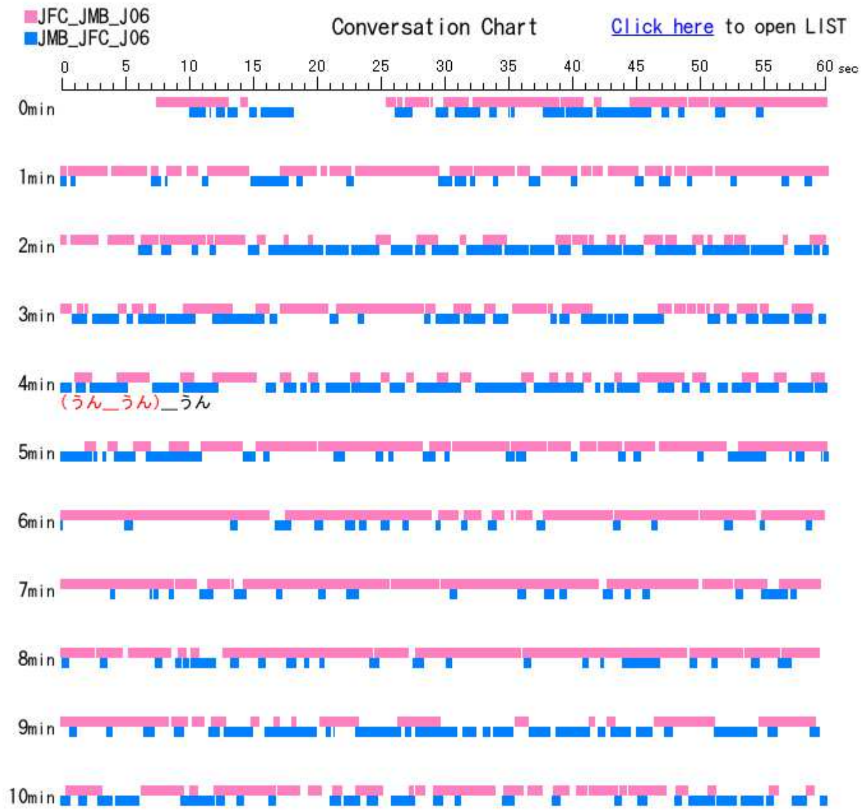

Figure 1.1. Speech & silence plots for the first 11 minutes of conversation #6 between two Japanese speakers, JFC and JMB, showing fragmentation of the discourse and progressive but not absolute alternations of speaker dominance. Each line shows one minute of speech, with speaker JFC's speech activity plotted above and that of speaker JMB plotted below. White space indicates lack of speech activity. The figure is taken from a screen capture of an interactive web page

very strained. The partners have to wait for their turn to speak, and there are long silences in the conversation.

A naturally interactive dialogue is not like a tennis match, where there is only one ball that can only be in one half of the court at any given time. Rather it is like a volley of balls being thrown in several directions at once. The speaker does not usually wait silently while the listener parses and reacts to an utterance; there is a constant exchange of speech and gesture, resulting in a gradual process of mutual understanding wherein a 'meeting of the minds' can take place **?**.

Natural Interactive Speech

As part of the JST/CREST Expressive Speech Processing (ESP) project **?**, we recorded a series of conversations between ten people who were not initially familiar with each other, and who had little or no face-to-face contact, but who were paid to meet once a week to talk to each other over the telephone for thirty-minutes each, over a period of ten weeks. The content of the conversations was completely unconstrained and left up to the initiative of the participants.

The volunteer speakers were paired with each other as shown in Figure 2 so that each conversed with a different combination of partners to maximise the different types of expressiveness in the dialogues without placing the speakers under any requirement to self-monitor their speech or to produce different speaking styles "on-demand". The ten speakers were all recorded in Osaka, Japan, and all conversations were in Japanese. Since the speakers were not familiar with each other initially, little use was made of the local dialect and conversations were largely carried out in the so-called 'standard' Japanese. Again, no constraints on types of language use were imposed, since the goal of this data collection was to observe the types of speech and the variety of speaking styles that 'normal' people use in different everyday conversational situations.

```
 female         male

( CFA EFA   CMA EMA )     (foreign)
    /            \          Group A
   JFA     -     JMA
    |             |
   JFB           JMB        Group B (baseline)
    |             |
   JFC     -     JMC
    |             |         Group C
  (Fam)         (Fam)      (intimate)
```

Figure 1.2. Showing the form of interactions between participants in the ESP_C corpus. The first letter of each participant identifier indicates the mother-tongue (Japanese/Chinese/English) of the speaker, the second letter indicates the speaker's sex (female or male), and the third letter is the group identifier. (Fam) is short for 'family'; indicating intimate conversations with relatives.

Figure 1 shows the speech activity patterns of two Japanese speakers, one male (JMB) and one female (JFC) for the first eleven minutes of their sixth thirty-minute telephone conversation. We can see that even

Table 1.1. Showing quantiles of speech activity time per speaker. ‘Silence’ is when neither is speaking, ‘overlap’ when both are speaking at the same time. ‘Sil’ shows the time each speaker individually (A or B) was quiet. ‘Solo’ shows the total duration of non-overlapping speech per speaker, and ‘talk’ the total overall speech time including overlaps. ‘Duration’ shows timing statistics for the entire conversation (assumed to be 30 minutes by default). All times are shown in minutes. Data are calculated from the time-aligned transcriptions of 100 30-minute conversations

	min	25%	median	75%	max
silence	0.99	2.08	2.85	3.81	7.03
silA	6.73	10.68	14.02	16.91	22.46
silB	5.72	13.09	14.68	17.68	21.58
soloA	4.14	9.51	11.66	14.68	18.17
soloB	4.55	8.39	10.64	13.32	18.90
overlap	2.66	5.53	7.01	9.04	12.80
talkA	10.80	16.04	18.75	22.44	28.52
talkB	12.20	15.66	17.93	20.15	27.15
duration	28.57	32.00	32.93	33.96	37.98

though it is usually quite clear who is the dominant speaker at any point in the conversation, neither speaker stays quiet for long, and that a gap of even 5 seconds in the speech could be considered as a long pause. In this example, the female speaker was older than the male and she tended to lead the conversation.

Table 1 gives details of speech activity time per speaker. It shows that for a 30-minute conversation between two people, median speaking time is approximately 18 minutes per speaker. There is approximately 3 minutes when no-one is speaking (10% of the total time) and 7 minutes (i.e., more than 20% of the conversation) when both speakers are speaking at once. Since time of non-overlapping speech is approximately 14 minutes per speaker, we can conclude that people overlap their speech, or talk simultaneously, one third of the time. These data were calculated from time-aligned transcriptions of 100 telephone conversations.

If we compare this ‘natural’ form of speech activity to that required for use of a speech translation system, we find that the waiting time imposed by the ‘tennis-type’ of speech interaction assumed in the technology is excessively long.

Two-way Interactive Speech

The controlled speech of professionals, such as broadcasters, newsreaders and announcers, is typically much closer to written text in form, since they are (a) usually practised and rehearsed, and (b) remote from their listeners. The speech of two people in face-to-face or telephone-based interaction, on the other hand, is neither practiced nor remote. The interaction requires a to-and-fro of information as the listener confirms, questions, and embellishes the speaker's propositional fragments. Being a two-way interactive process, it also requires some form of discourse management control.

It is common to speak of disfluencies in natural speech, and of fillers and hesitations as if they are performance errors, with the assumption that 'perfect' speech would be very similar in form to a written text, like that of a professional, well-formed, clear, concise, and precise. However, we might also consider an alternative point-of-view, as proposed here, that this so-called 'ill-formed' speech is in fact the product of natural evolution of the spoken language so that it can transmit interpersonal, affective, and discourse-related information *at the same time as,* and *in parallel to,* the transmission of propositional content.

To account for this supposedly 'broken' form of natural conversational speech we have suggested a structure of of 'wrappers' and 'fillers' **?** wherein the propositional content, here called a 'filler' (the term is used here as if describing the contents of a box of chocolates, with each wrapped distinctively and all having different fillers) is 'wrapped' in affect-bearing prefix and suffix fragments.

In this hypothesis, the speaker forms a complex utterance through a sequence of smaller and simple fragments. These are not presented in a concise linear sequence as they might be in writing, but are interspersed with 'wrappers' that indicate how they should be perceived. The speaker typically has a large repertoire of semantically 'empty' but affectively marked words or phrases (such as fillers in the conventional sense) that can be added at the beginning or end of an utterance fragment to embellish it or to show affect-related information. Some examples for Japanese, with their counts, are given in Table 3.

For an example in English, we can consider the speech of a typical Londoner who might produce the following sequence:

> *" ... erm, anyway, you know what I mean, ..., it's like, er, sort of* **a stream of** *... er ...* **words, and** *you know* **phrases***, all* **strung together***, you know what I mean, ... "*

where the words in bold-font form the content (or the *filling* of the utterance) and the italicised words form the *wrapping* or decoration around the content.

This (mis-)usage of the term filler is in (deliberate) contrast to its usual interpretation as something which 'occupies a gap' or a supposed empty space in a discourse. On the contrary, we suggest here that these are not gaps in the discourse but essential markers for a parallel tier of information. By their very frequency, these non-propositional and often non-verbal speech sounds provide not just time for processing the linguistic content of the spoken utterance but also a regular base for the comparison of fluctuations in voice-quality and speaking-style that indicate how the content is to be understood and how it relates to the discourse.

These fragments allow the speaker to express information related to mood and emotion, to interpersonal stance, and to discourse management. By being effectively transparent (i.e., they would not be transcribed when recording the speech in the minutes of a meeting, for example) they do not interfere with the transmission of linguistic or propositional content, but by being simple, frequent, and often repeated sounds, they allow easy comparision, like-with-like throughout the utterance so that the listener can sense the speaker's intentions through subtle variation in their usage and prosody.

Speech Fragments

From an analysis of 150,000 transcribed conversational utterances in a separate section of the JST-CREST ESP corpus, recorded from one female speaker over a period of four years, we found that almost 50%of the utterances are 'non-lexical'; i.e., they could not be adequately understood from their text alone. (Table 2 provides detailed figures, Table 3 shows some examples). Very few of these utterance types can be found as an entry in a standard language dictionary, yet it was confirmed that the intended meanings of many of these non-verbal utterances (or conversational 'grunts') can be perceived consistently by listeners even when presented in isolation without any discourse context information. In many cases, the intentions underlying the utterances can be appropriately and consistently paraphrased even by listeners of completely different cultural and linguistic backgrounds **?**.

Table 1.2. Counts of non-verbal utterances in the transcriptions for one female speaker in the ESP corpus. Utterances labelled 'non-lexical' consist mainly of sound sequences and combinations not found in the dictionary, but may also include common words such as "yeah" "oh", "uhuh", etc.

total number of utterances transcribed	148772
number of unique 'lexical' utterances	75242
number of 'non-lexical' utterances	73480
number of 'non-lexical' utterance types	4492
proportion of 'non-lexical' utterances	49.4%

Table 1.3. Counts of the hundred most common utterances of Japanese, as found in the ESP corpus of natural conversations. All function to display affect. While direct glosses are not provided here, most would be transcribed as variants of ummm, aah, uhuh, yeahyeahyeah, etc., @S, @E, and @K are symbols used to indicate breath-related sounds such as a hiss or a sharp intake of breath to show surprise or displeasure.

10073	うん	467	ズーー	228	ううん	134	へー.ーー
9692	@S	455	スー	227	えっ	134	はい.はい.はい.はい
8607	はい	450	んーー	226	へ.ーー	134	そう.です
4216	laugh	446	うーーーん	226	ハハハ	133	@E
3487	うーん	396	ねー	225	う.んー	133	あ.そう.な.ん.です.か
2906	ええ	395	あ.あー	200	そうですね	130	そう.な.ん.です.か
1702	はーい	393	はい.はい.はい	199	ほ.ーー	129	は.ー
1573	うーーん	387	あー.はい	193	ハー	129	い
1348	ズー	372	ねえ	192	その	127	ほ.ー
1139	ふん	369	ふーーん	190	え.えー	125	ハハハハハ
1098	あのー	369	だから	188	あ.あーー	119	はい.はい
1084	あっ	368	あー.ん	187	ね	119	は.ーーー
981	はあい	366	ああ	180	ん.はい	114	ハハ
942	あの	345	あの.ーー	180	あの.ーーー	113	は
941	ふーん	337	なんか	173	ん.ん	113	で.ー
910	そう	335	え	172	アハハハ	113	て
749	えー	311	でも	168	はい.ー	112	は.あー
714	あーー	305	スーー	164	う.うーん	110	フフフ
701	あ	274	うん.うん.うん	161	は.ーー	110	そのー
630	あーーー	266	ハハハハ	160	@K	110	もう
613	あ.はい	266	て.ー	159	そう.です.ねー	109	ふーーーん
592	うん.うん	266	え.ーー	151	あーーーー	108	はあ.ーー
555	あー	258	で	143	だから.ー	106	そうですね.え
500	んー	248	う	139	アハハハハ	105	んー.ん
469	ん	242	へー	137	そう.そう.そう	104	いや

In the following subsection, we will see how these affect-bearing fragments, which are effectively transparent in the discourse and do not appear at all intrusive to an observer, can carry significant interpersonal information through tone-of-voice and other such prosodic variation.

2. Acoustic Correlates of Discourse-related Non-verbal Speech Sounds

In previous work we have found from an analysis of the speech of a single female speaker that her voice quality changes significantly according to type of interlocutor, familiarity with the interolocutor, pragmatic force of the utterance, etc. In this paper we add further evidence to show that this is a general phenomenon, using speech data taken from a series of recorded telephone conversations between a small number Japanese men and women over a period of several months.

Voice-quality, Prosody, and Affect

The study based on analysis of the ESP corpus of conversational-speech **?** showed that voice quality, or laryngeal phonation style, varied consistently and in much the same way as (but independently of) fundamental frequency, to signal paralinguistic information **?**. We showed that the factors 'interlocutor', 'politeness', and 'speech-act' all had significant interactions with this variation.

The mode of laryngeal phonation can be measured from an estimate of the glottal speech waveform derivative (a result of inverse filtering of the speech using time-varying optimised formants to remove vocal tract influences **?**) by calculating the ratio of the largest peak-to-peak amplitude and the largest amplitude of the cycle-to-cycle minimum derivative **?**. In its raw form it is weakly correlated with the fundamental period of the speech waveform ($r = -0.406$), but this can be greatly reduced by $NAQ = log(AQ) + log(F_0)$, yielding a Normalised Amplitude Quotient (henceforth 'NAQ') **?** ($r = 0.182$).

We analysed data from one female Japanese speaker, who wore a small head-mounted, studio-quality microphone and recorded her day-to-day spoken interactions onto a MiniDisk **?**, **?**, **?** over a period of more than two years. The data comprise 13,604 utterances, being the subset of the speech for which we had satisfactory acoustic and perceptual labels. Here, an 'utterance' is loosely defined as the shortest section of speech having no audible break, and perhaps best corresponds to an 'intonational phrase'. They vary in length from a single syllable to a thirty-five-syllable stretch of speech.

The factor 'interlocutor' was analysed for NAQ and F_0, grouped into the following classes: Child (n=139), Family (n=3623), Friends (n=9044) Others (n=632), and Self (n=116). It is clear that F_0 and breathiness are being controlled independently for each class of interlocutor. Repeated t-tests confirm all but the child-directed (n=139) voice-quality differences to be highly significant.

Figure 2 (left part) shows median NAQ and F_0 for the five categories of interlocutor. The values are z-scores, representing difference from the mean in SD units. NAQ is highest (i.e, the voice is breathiest) when addressing 'others' (talking politely), and second highest when talking to children (softly). Self-directed speech shows the lowest values for NAQ, and speech with family members exhibits a higher degree of breathiness (i.e., it is softer) than that with friends. F_0 is highest for child-directed

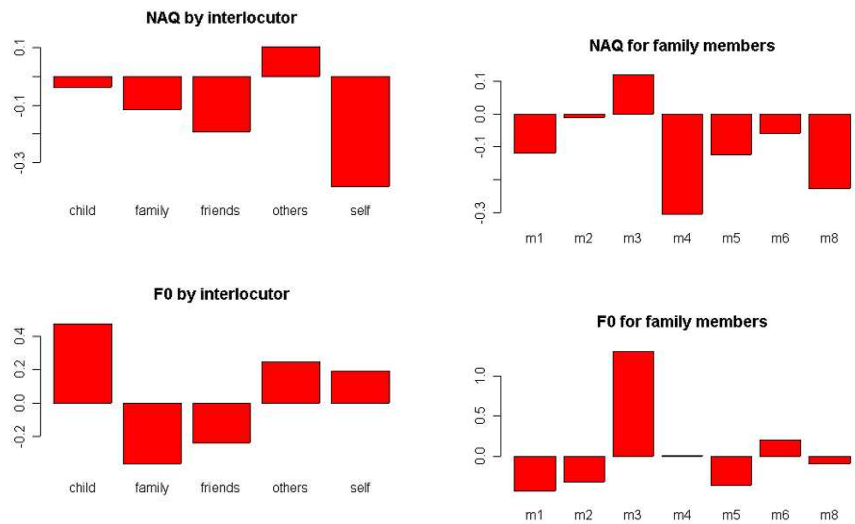

Figure 1.3. Median values of NAQ and F_0 plotted for interlocutor (left) and for family members (right). m1: mother, m2: father, m3: daughter, m4: husband, m5: older sister, m6: sister's son, m8: aunt. Data are (z-score) scaled, so values are in SD units. 0 represents the mean of the distribution

speech, and lowest for speech with family members (excluding children). Figure 2 (right part) shows the values for 'family' speech in more detail. It reveals some very interesting tendencies. Family members can be ordered according to breathiness as follows: $daughter > father > nephew > mother = older\ sister > aunt > husband$. Thus, it seems that the ordering reflects the degree of 'care' taken in the speech to each family member. In traditional Japanese families, the father is perhaps a slightly remote figure, but deserves respect. The mother (and older sister) come next in ranking, and husband comes last - not indicating a lack of respect, but an almost total lack of need to display it in the speech. We can see from the data here that this speaker also has a very close relationship with her aunt, a detail that was confirmed by her in person later.

Multi-speaker Variation in Prosody and Tone of Voice

To further validate this finding, we recently processed the data from the ESP_C corpus of telephone conversations between people who were strangers at first but then gradually became friends over the period of the recordings. These Japanese adults used head-mounted microphones and recorded their speech directly to DAT while they spoke to each other over the telephone from different locations with no face-to-face contact.

At the beginning of the recordings, they were all strangers to each other, but over the period of ten weekly conversations they gradually became familiar to differing degrees. They spoke over the telephone to each other, to family members, and to foreign visitors to Japan who were capable of holding a conversation but not fluent in the language. In this way, we were able to control for 'ease of communication' without constraining the conversations in any artificial way. They were paid to talk to each other and, from the transcriptions of the dialogues, appeared to enjoy doing so.

Because the calculation of NAQ requires a degree of hand intervention for setting up the initial speaker-related parameters, we decided to use a combination of several measures of prosodic information that could all be extracted automatically, without manual intervention, from the speech waveform. We extracted acoustic data from the recordings of both speakers in a series of 100 30-minute conversations.

A combination of 14 different acoustic features was used in this experiment. The mean, maximum, minimum of power (rms amplitude) and pitch (F_0), the position of the F_0 peak of each utterance, measured as a percentage distance from 0 (beginning) to 100 (end of utterance), the amount of voicing throughout the utterance, the values of the first and second harmonics, the third formant, and the spectral tilt (after Hanson **?**), as well as a measure of speaking rate or normalised duration of the utterance. These measures were averaged across the whole of each utterance, giving only a general indication of prosodic settings for longer utterances but allowing a very precise comparison of the more frequent shorter utterances when comparing like with like throughout the progress of a discourse.

We performed a Principal Component Analysis (pca) of these data to reduce the number of factors in the measure, and then plotted the first three principal components, which account for about half of the variance observed in the acoustic data, categorised by conversation number. In this way we can show how the prosodic settings vary with time. In the default case we would expect them to remain the same over time. For example, a person's voice pitch may change a little from day to day, according to health, smoking, and alcohol intake, as well as according to mood and emotion, but we would expect to see a steady average over a period of several weeks.

Table 1.4. Results of the Principal Component Analysis

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.65	1.53	1.38	1.32	1.12	0.96	0.89	0.83
Proportion of Variance	0.19	0.16	0.13	0.12	0.08	0.06	0.05	0.04
Cumulative Proportion	0.19	0.36	0.49	0.62	0.71	0.78	0.83	0.88

	PC9	PC10	PC11	PC12	PC13	PC14
Standard deviation	0.74	0.71	0.61	0.29	0.23	0.0004
Proportion of Variance	0.03	0.03	0.026	0.006	0.004	0.0001
Cumulative Proportion	0.92	0.96	0.98	0.99	1.00	1.00

Table 4 gives details of the pca analysis, showing how much of the variation was covered by each component. Table 5 shows how the individual acoustic measures were mapped by the components in the pca reduction. We can see that approximately half of the variance is covered by the first thee components alone, and that more than 80% is accounted for by the first seven.

From Table 5 we can see that the first principal component maps well onto F_0 mean and maximum, while the second maps onto h1 (power at the first harmonic) and h1a3 (the ratio of first harmonic to amplitude of the third formant) which is a measure of spectral tilt related to breathiness and tension in the voice. The third component has a broader scope but appears related to degree of voicing and measures of signal amplitude.

It is encouraging that these automatically derived measures match well to our intuitions mentioned above about the usefulness of measures of spectral tilt as a prosodic feature. At the interpersonal level of spoken interaction, tone-of-voice is perhaps more important than e.g., pitch patterns, which form the core of traditional prosodic research and have a closer relation to syntactic and semantic structures within the linguistic component of the utterance.

Also of great interest is the finding shown in Figure 4 is that the first three components (at least) vary in a consistent way with progression of the conversations through the series. We can see a clear increase in values of each component, going from negative in the earlier conversations to positive in the later ones. This correlates well with the increase in familiarity between the participants and shows that their basic phonatory settings change. The discrepancy seen in the final conversation may well arise as a result of that conversation being recorded (as an afterthought)

Table 1.5. Showing the precise relationship between each principal component and each prosodic factor derived automatically from the acoustic speech signal

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13	PC14
fmean	-50	14	25	-20	11	-3	4	-7	8	3	14	27	71	0
fmax	-48	10	23	0	9	1	30	-26	18	18	36	-37	-46	0
fmin	-29	11	17	-46	12	-26	-27	15	7	-14	-62	-9	-26	0
fpct	-6	20	-13	-22	-40	46	-62	-25	11	18	12	-9	-2	0
fvcd	-8	-23	-39	-29	-7	6	36	-61	-15	9	-28	29	-7	0
pmean	-36	-26	-31	29	-16	-7	-7	-7	-17	-28	-16	-59	30	0
pmax	-43	-12	0	42	-5	2	-27	1	-17	-30	8	56	-34	0
pmin	-20	-26	-37	7	-12	-34	-10	32	29	64	2	11	-1	0
ppct	-16	16	-15	-14	-51	30	45	46	23	-27	-6	8	-5	0
h1h2	-13	-30	32	4	6	53	8	25	-44	41	-26	-5	1	0
h1a3	9	-50	34	-12	-28	-9	-4	-9	20	-12	7	1	1	-67
h1	5	-57	11	-14	14	19	-7	-1	43	-21	8	0	0	60
a3	-8	0	-39	-1	63	40	-5	12	27	-10	0	-1	-1	-44
dn	5	18	22	54	-6	12	7	-27	49	14	-51	5	9	0

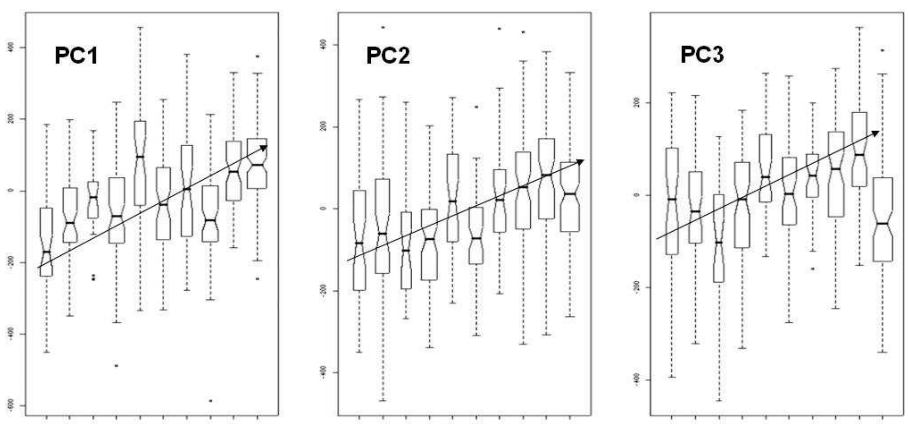

Figure 1.4. The fiast three principal components plotted by conversation number for speaker JMC. We can see clear indication of an increasing trend that correlates with familiarity of the participants through the series of conversations.

after a longer break, to make up for a missed conversation earlier in the series.

3. Technological Applications

At this point we should perhaps consider how these findings can be made use of in speech technology applications. We can immediiately consider two aspects of future development: one concerned with discourse flow, the other with affect sensing. The first will allow people easier access to machine-mediated speech; the second will allow machines access to people-related information that may not be immediately discernible from the linguistic output of a speech recogniser.

Discourse Flow and Prosody Engineering

Currently, the users of a speech translation system have to wait in silence until their utterance and its reply have both been translated. We have seen from the above data that this form of interaction is quite unlike that of normal human-human discourse.

Just as aeroplanes do not flap their wings when they fly, it may be the case that this is the optimal mode of usage for such a device, and that emulation of human speech habits may be inappropriate in such a technology. However, it might on the other hand feel more natural for a user if the machine gave encouraging feedback while the conversation was in progress, or if there was some mechanism for the speaker to communicate in fragments rather than in well-formed whole sentences.

Since the machine often has some knowledge of its domain, whether through ontologies, dictionaries, or example corpora, it should be feasible to generate a dialogue interactively by the mutual exchange of fragments. As in a human conversation, where both partners echo, repeat, check, suggest, and challenge their mutual understanding of the present state of the dialogue, so a prosody sensor should be able to use tone-of-voice information in addition to the recognised text input to add fragments appropriately onto the 'understanding stack'.

Since the number of wrapper-type fragments is small (on the order of a hundred or so) they can easily be stored as a dictionary. For each entry a further set of codebook entries detailing the acoustic characteristics of the common prosodic and voice-quality variants can then be stored as a sub-dictionary listing. We have found a codeboook size of 16 to be optimal here. As each is recognised, by simple pattern matching, its subvariant is selected and a flag indicating supposed speaker intention & state added to the discourse stack. Integrating this component into a translation system, however, remains as work in progress.

Sensing Affect; Detecting Changes in People from Variation in their Speaking Style and Tone of Voice

There is increasing interest nowadays in the areas related to Affective Computing (see e.g., **?**, **?**), particularly with respect to sensing human states and conditions from external physical cues. Since it is likely that people sense and respond intuitively to the small affect-related changes in prosodic settings when conversing with a human partner, and it would be socially beneficial if a machine could also be made sensitive to these cues from the voice.

There has recently been a call in Japan **?** for research into proactive devices for use in an advanced media society. Currently, most mechanical devices work *reactively*, responding to a command from a user, but certain funding agencies here are hoping that future machines will be able to anticipate the user requirements and perform an appropriate function *proactively*, without explicit prior control from the user. For these technologies, a degree of quite sophisticated human sensing will be required. However, although the technology itself will be very sophisticated, the information that is being sensed may be quite low-level and primitive.

For example, in the meetings-related research (**?**, **?**, **?**, **?**) sensor devices have been invented that detect different degrees of human participation in a multi-party dialogue from simple cues such as amount of bodily movement and coincidences in the timing of simple actions such as nodding. Similar cues can be detected from tone-of voice, laughter, and non-verbal speech sounds that are currently regarded as insignificant.

Machines can be trained to produce a given response when more than one person laughs or when even one person makes a given sound (such as a disapproving grunt). By processing differences in the timing, prosody, or frequency of these cues, much information can be gained into the mental states and intentions of the participants.

Towards the Synthesis of Expressive Speech

If we are to produce speech synthesis that resembles human speech in a conversational setting, then we will need a grammar of such non-verbal utterances, and language models that predict how often, when, and which non-verbal speech sounds should be generated in a discourse. Much of this remains as future work, though several proposals have already been made (see e.g., ieee). Because they are typically short single utterances (or discrete groups of repeated syllables), there is no need to calculate a join cost for concatenative synthesis, and these non-verbal speech sounds can be inserted easily into a stream of synthesised speech.

However, because their target prosody can vary not just in pitch but also in voice-quality, there is need for a precise and finely-tuned target cost instead.

There is already considerable research being carried out into the generation of synthetic speech with emotion, but little into the generation of speech with noticeable differences in speaker intention. Interestingly, in our analysis of five years of conversational speech recorded in ordinary everyday environments, the amount that was markedly emotional accounts for less than 1% of the total; the amount that is marked for affect is probably more than half.

This difference may be a result of volunteers hesitating to give us speech that was openly emotional. If they happened to have a blazing argument with their partner on a given day, for example, they may have deleted the recording out of embarrassment or a sense of privacy. Yet the amount of potentially embarrassing personal information that they *did* give us, without hesitation, leads the author to believe that this is not the case. It is more likely that as socially-responsible adults, we moderate our speech so as not to reveal personal emotional details most of the time. We use it instead to show interest, enthusiasm, boredom, concern, care, relief, etc., i.e., *to appear* bright, cheerful, intelligent, etc., rather than to *reveal* our actual inner feelings and emotions.

Since there is already an exhaustive literature on the relations between prosody and syntactic structure, prosody and semantics, and the use of prosody in the expression of contrastive focus, etc., we will not address those issues further here, but instead we claim that the role of affective prosody in interactive speech is more to show the partner the speaker's intentions, to clarify stages of the discourse, and to manage turn-taking. This functional interpersonal role of prosody leaves plenty of scope for future research.

There are many applications, apart from human-to-human speech translation, where a natural-sounding voice is required in speech synthesis. This paper has shown that for the voice to be completely natural-sounding, a new level of language structure and discourse control will need to be incorporated into future speech synthesis research.

4. Discussion

The author, being of the UNIX persuasion since the early eighties, has recently been obliged to make use of the Windows operating system due to the needs of publishers and conference organisers. Since disk access can sometimes be very slow when the files are fragmented under this operating system, one soon learns the benefit of frequent use of two MS-

Dos commands: Disk-Cleanup and Disk-Defrag. While first considering this as a design fault in the operating system itself, I now consider that it may indeed represent the 'natural way of things'. Natural speech appears to be similarly fragmented. It appears that when we listen to natural interactive or conversational speech we also perform considerable clean-up, to remove hesitations and 'wrappers', and then defrag the segments to produce intelligible chunks from the speech sequence.

Accordingly, we suggest in this paper that the evolution of this supposedly 'broken' form of spontaneous speech is not just a side-effect of poor performance in real-time speech generation processes, but that the inclusion of frequently repeated non-verbal speech segments allows the speaker to use them as carriers for affective information such as is signalled by differences in voice quality and speech prosody. Their high frequency (and relative transparency with respect to the propositional content) allows small changes or contrasts in phonation style to be readily perceived by the listener, even if he or she is at first unfamiliar with the speaker.

5. Conclusion

This paper has presented some acoustic findings related to speech prosody and has shown how the voice is used to signal affective information in normal conversational speech.

The paper has shown that natural conversational speech is highly fragmented, that these fragments carry discoursal and affect-related information, and that by being very frequent, and effectively transparent to the discourse, they function as an efficient carrier for this second channel of information in interactive speech.

The paper has argued that although modern speech processing technology has come a long way, and appears now to have achieved many of its original goals, it is perhaps now time to 'shift the goalposts' and become more aware of this secondary channel of information carried in the speech signal that is currently not being processed as part of the human communication process.

Acknowledgements

This work is partly supported by the Ministry of Public Management, Home Affairs, Posts and Telecommunications, Japan under the SCOPE funding initiative. The ESP corpus was collected over a period of five years with support from the Japan Science & Technology Corporation (JST/CREST) Core Research for Evolutional Science & Technology funding initiative. The paper was written while the author was employed by the National Institute of Information and Communications Technology. The author also wishes to thank the management of the Spoken Language Communication Research Laboratory and the Advanced Telecommunications Research Institute International for their continuing support and encouragement of this work.

References

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

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

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,

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.

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

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

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.

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.

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)

Anderson, Terry L., and Fred S. McChesney. (n.d.). "Raid or Trade? An Economic Model of Indian-WhiteRelations," Political Economy Research Center Working Paper 93–1.

Lacey, W.K. (1968). *History of Socialism.* Ithaca, NY: Cornell University Press.

Oliva, Pavel. (1971). *Sparta and Her Social Problems.* Amsterdam: Adolf M. Hakkert.

Zimmern, Alfred. (1961). *The Greek Commonwealth: Politics and Economics in Fifth-Century Athens,* 5th ed. New York: Galaxy Book, Oxford University Press.