How Speech Encodes Affect and Discourse Information

Conversational Gestures

Nick Campbell

[1] *National Institute of Information and Communications Technology*
[2] *ATR Spoken Language Communication Research Laboratory,*
Keihanna Science City, Kyoto 619-0288, Japan.
nick@nict.go.jp & nick@atr.jp

Abstract. This paper presents an analysis of several recorded conversations and shows that dialogue utterances can be categorised into two main types: (a) those whose primary function is to impart novel information, or propositional content, and (b) those whose primary function is to relay discourse-related and interpersonal or affect-related information. Whereas the former have characteristics that are closer to read speech, the latter are more varying in their prosody and present a considerable challenge to current speech synthesis systems. The paper shows that these apparently simple utterances are both very frequent and very variable, and illustrates with examples why they present such a difficult challenge to current speech processing methods and synthesis techniques.

Keywords. Speech Technology, Discourse, Affect, Non-verbal Speech Communication

Introduction

It is well known that "The act of sending and receiving messages is a process of negotiation of meaning wherein both the sender and the receiver are mutually responsible for the creation of this meaning" [1].

In the context of speech technology, there are already well-developed devices for the recognition of verbal speech and for the processing of its propositional content, but very little is yet known about methods for processing the non-verbal content in a dialogue speech signal. We can process broadcast news easily, but still perform poorly when faced with ordinary dialogue speech.

This paper describes some current work towards the processing of non-verbal speech utterances in a dialogue context. The work is based upon an analysis of a very large corpus of everyday spoken interactions captured under extremely natural situations [2]. The paper presents a view of speech interaction as not only facilitating the exchange of linguistic or propositional information, but also facilitating the display of affect, interpersonal stances, and social relationships. By incorporating such non-verbal content in a model of speech communication it may become easier to model the subtle two-way inter-

actions between both speaker and listener that are necessary for facilitating the transfer of meaningful discourse content.

Concurrent work [3] using the same model is being carried out towards the production of a 'conversational' speech synthesis system for use in interactive dialogues, such as might take place between a person and an information system, a robot, or a speech translation device. There are several types of response and feedback utterances that are currently very difficult to implement using traditional speech synthesis methods, yet these non-verbal speech sounds or 'conversational gestures' function to provide status-updates in an interactive discourse. Such (often phatic) utterances include laughter and grunts as well as many common phrases and idioms, and their choice and variety can reveal much information about the speaker's (i.e., the current listener's) states in an interactive discourse.

This model of information exchange incorporating non-verbal backchannel speech utterances shows how feedback from the listener is used to help the speaker both to deliver content more efficiently, and at the same time to be reassured of the degrees of success in the flow of information transmission. It assumes that both the sender and the receiver are equally responsible for the mutual creation of meaning for each segment in a discourse and that they do this through the mediation of non-verbal cues. This paper will concentrate on those cues that are signalled by speech, which we refer to here as the 'audio landscape'.

1. The Audio Landscape

The 'audio-landscape' of a discourse enables a participant or observer to estimate the types of interaction and to make guesses about the relative status of participants without the need for a content-based analysis of any given utterance or sequence of utterances. In other words, even a foreigner who has no understanding about the specifics of what is being said can often make an intelligent guess about the functional states, i.e., about what is happening in a dialogue at the interpersonal level.

By simply watching what is happening in a conversation, without even any sound information at all, we can often see who is doing what; not just who is speaking (which can be determined relatively easily from the amount of bodily movement, for example), but also who is listening (which can be determined from the synchrony of movements related to events in the speech) as has been shown by e.g., the early work of Kendon & Condon [4,5] and the more recent 'meetings' findings [6,7,8].

Furthermore, if in addition to the visual information we also have access to the sound, then we can make an intelligent guess about how the participant listeners are reacting to the content of each utterance, even if (like the foreigner) we do not understand the content of the speech itself. Laughs, nods, grunts, and other such speech gestures serve to indicate the degrees to which the listener is attentive, synchronised with the content of the discourse, and in relative states of agreement with it. This much can be determined from the non-verbal content [9,10].

We are currently performing research into technology to process this audio landscape in order to detect the main speaker in a given discourse situation, both in a meeting environment [11] and in general two-person conversations, to categorise the competing forms of speech in a given situation. Several speech gestures such as laughter, agreement,

and feedback-responses can be recognised, isolated, and used to determine the progress of the meeting and the degrees and types of participation status among the members present.

2. Data Collection

As part of the JST/CREST Expressive Speech Processing (ESP) project [12], a series of conversations were recorded 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. These recordings constitute the ESP_C subset of the ESP corpus.

The volunteer speakers were paired 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 used in different everyday conversational situations.

Four of the ten speakers were non-native; their inclusion was not so that we should have foreign-accented speech data, but rather that we should be able to observe changes in the speech habits of the Japanese native speakers when confronted with linguistically-impaired partners. Two were male, two female, two Chinese, and two English-language mother-tongue speakers. These and the two Japanese who spoke with them formed Group A in our study. Group B is the 'baseline' group, consisting of a male and a female Japanese native speaker who conversed in turn with the each other and with the Japanese native speakers of both sexes from Groups A and C. Group C similarly consisted of a

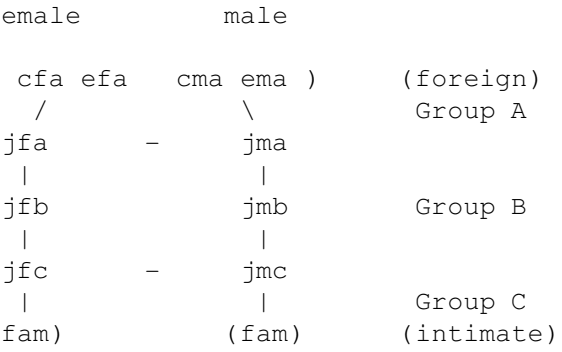

Figure 1. Showing the form of interactions between participants in the ESP_C corpus. The first letter of each participant identifier indicates the mother-tongue (Japanese/Chinese/English) of the speaker, the second letter indicates the speaker's sex (female or male), and the third letter is the group identifier. (fam) is short for family; indicating intimate conversations with relatives.

```
CFA JFA C01 200.369 0.491 #
CFA JFA C01 200.860 0.808 laugh
CFA JFA C01 201.668 0.869 あとは
CFA JFA C01 202.537 1.099 変わりました
CFA JFA C01 203.636 1.868 laugh
CFA JFA C01 205.504 0.670 うん
CFA JFA C01 206.174 0.744 #
CFA JFA C01 206.918 0.917 はい
CFA JFA C01 207.835 2.691 #
CFA JFA C01 210.526 0.602 はい
CFA JFA C01 211.128 2.791 #
CFA JFA C01 213.919 0.749 @S
CFA JFA C01 214.668 2.685 そうです結構もう四年間なりました
CFA JFA C01 217.353 0.785 はい
CFA JFA C01 218.138 0.561 #
CFA JFA C01 218.699 0.731 はい
CFA JFA C01 219.430 1.384 #
CFA JFA C01 220.814 1.088 行ってます
CFA JFA C01 221.902 0.738 #
CFA JFA C01 222.640 0.784 はい
CFA JFA C01 223.424 1.107 #
CFA JFA C01 224.531 1.356 あの一歳です
CFA JFA C01 225.887 0.525 #
CFA JFA C01 226.412 0.600 はい
CFA JFA C01 227.012 2.795 #
CFA JFA C01 229.807 0.443 はい
CFA JFA C01 230.250 0.941 #
```

Figure 2. Transcription was performed by hand, using the Transcriber software package. The first 3 columns in the figure identify the speaker, partner, and conversation number. The numbers represent the start time of each utterance in the conversation (in seconds) and its duration. Laughs, non-speech noises, and silences are also transcribed along with the text.

male and a female Japanese native speaker who conversed with each other and with the members of Group B, but who also telephoned their own family members each week and spoke with them for a similar amount of time. Figure 1 illustrates these pairings graphically.

The corpus thus obtained allows us to examine the prosodic characteristics and speaking habits of Japanese native speakers when confronted with a range of different partners on the spectrum of familiarity, and to observe changes in their speech as this familiarity changes over time. Our principal targets for this series of recordings were the six Japanese native speakers (three male and three female) who came to an office building in Osaka once a week to answer the telephone and speak with each partner for a fixed period of thirty-minutes each time. All wore close-talking, head-mounted, Sennheiser microphones and recordings were taken directly to DAT with a sampling rate of 48kHz. The offices were air-conditioned, but the rooms were large and quiet, and no unwanted

```
JFA CFA C01 203.276 1.362 4456 ==> <[ laugh ]>
JFA CFA C01 204.638 0.902 0 ==> <[ @S ]>
JFA CFA C01 205.540 1.927 0 +->  <[あーそうなんですか]>
JFA CFA C01 207.467 0.322 0 ==> <[ @S ]>
JFA CFA C01 207.789 0.401 0 ==> <[ はい ]>
JFA CFA C01 208.190 0.227 0 ==> <[ @S ]>
JFA CFA C01 208.417 1.744 0 ==> <[ あのー ]>
JFA CFA C01 210.976 0.393 814 ==> <[ え ]>
JFA CFA C01 211.369 0.260 0 ==> <[ え ]>
JFA CFA C01 211.629 1.139 0 --> お,ご結婚をきょ
JFA CFA C01 212.768 0.264 0 ==> <[ え ]>
JFA CFA C01 213.032 1.566 0 --> 何時なさったとおっしゃいまし
JFA CFA C01 216.356 0.687 1757 --> 四年目
JFA CFA C01 217.043 0.301 0 ==> <[ @S ]>
JFA CFA C01 217.344 1.498 0 +-> あ,<<そうです>>か
JFA CFA C01 218.842 0.422 0 ==> <[ @S ]>
JFA CFA C01 219.264 0.241 0 ==> <[ え ]>
JFA CFA C01 219.505 1.193 0 --> お子さんは
JFA CFA C01 221.686 0.283 987 --> X
JFA CFA C01 221.969 0.819 0 --> あ,いらっしゃる
JFA CFA C01 223.180 0.360 392 ==> <[ あ ]>
JFA CFA C01 223.540 1.248 0 --> お幾つです.か
JFA CFA C01 225.571 0.749 783 --> 一才
JFA CFA C01 226.320 0.347 0 ==> <[ @S ]>
JFA CFA C01 226.667 1.235 0 --> あっそう,じゃ
JFA CFA C01 227.902 1.891 0 +-> こういう<<ときは>>どういう風に
JFA CFA C01 229.793 1.494 0 +-> お子さんはされてる<<んですか>>
JFA CFA C01 231.287 0.746 0 ==> <[ @S ]>
JFA CFA C01 232.033 0.798 0 --> お家
JFA CFA C01 234.539 0.424 1707 ==> <[ あ ]>
```

Figure 3. Part of a dialogue, showing frequent utterances (n>=100) in < [square]> brackets, and frequent segments (N>=100) as part of longer utterances in < < angle> > brackets, which may be embedded. Speaker, listener, conversation number, start time and duration (in seconds) and delay (milliseconds) from end of previous utterance are also shown. "@S" indicates a sharp sucking intake of breath, a common speech gesture in Japanese. The paper argues that these very frequent interjections carry a separate stream of information through their prosody

noises (or acoustic reflections) were present in the recordings.

The speakers were all mature adults who were employed part-time by the recording agency and were paid for their participation in the recordings. They were initially unfamiliar with each other, but the degree of familiarity naturally increased throughout the period of the ten conversations. All have signed consent forms allowing the contents of the recordings to be used for scientific research. The ultimate purpose of the data collection was not made specific to the participants who were only told that their speech would be recorded for use in telecommunications research.

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

Table 1. The hundred most frequent single utterances in the ESP_C corpus. The numbers indicate the count of each word or phrase when it occurs as a single utterance in the transcriptions. Since duration is usually considered as distinctive in Japanese, the lengthening (an extra mora beat is indicated by a dash) may be significant. Note the highly repetitive nature of many of these utterances, very few of which can be found in any standard dictionary of Japanese. Note that these few samples alone account for more than a third (n=72,685) of the 200,000 utterances in the corpus. Less then half (n=92,541) of the utterances were unique.

3. Data Characteristics

Figure 2 shows part of a typical dialogue segment from Chinese speaker CFA, talking with her Japanese partner JFA during their first conversation. We can see even from this very short sample that there is considerable repetition; in this case of the word 'yes' (or its Japanese equivalent), interspersed with occasional longer content utterances. Table 1 lists the 100 most-frequent expressions from a corpus of 200,000 such dialogue utterances transcribed from recordings of the six people's telephone conversations. We (even those of us who cannot yet read Japanese) can see from this table that repetition is a common identifying characteristic of these frequently-repeated utterances. The same syllable (Japanese character or character sequence) repeats in more than half of the cases. If we expand this list to include the less frequent utterances, then we will find that they differ primarily in the number and type of repeats.

Among these repeats, we can discern several different patterns or types: one uses progressive lengthening (hah: はー, はーー, はーーー) (hmm: ふん, ふーん, ふーーん, ふーーーん), another simple repetition (ha!: ハ, ハハ, ハハハ, ハハハハ, ハハハハハ),

or complex repetition (umm: うん, うんうん, うんうんうん), and yet another increasing complexity(so: そうです, そうですね, そうですねー, そうなんですか, あそうなんですか). The hundred utterance types shown in the table above account for more than a third of the total number of utterances in the corpus. If we include their less frequent (typically longer) variants, then we find that more than half of the utterances in the corpus are of this non-verbal type (not usually found in a standard dictionary).

If we exclude these feedback utterances (i.e., just listen to those utterances marked with "–> " in Fig.3), then we can still understand the propositional part of the discourse, almost without change, but we lose the 'landscaping' information. Alternatively, if we just listen to those primarily non-verbal utterances ("==> " in Fig.1), then we can follow much of the interaction (in 'foreigner mode') without knowing anything about the content of the discourse. i.e., we can interpret the prosody to make an inference about the function of each utterance without knowing its specific lexical meaning.

3.1. Features of Non-Verbal Speech

Unlike regular lexical items which have a fixed form and a variable prosody depending on contextual information, these non-verbal 'speech gestures' rather seem to have a fixed prosodic identity (or underlying prosodic dynamic) and a variable form, extending to meet the requirements of the prosodic dynamics that they function to substantiate. Like bodily gestures, which have a few basic finite forms but considerable freedom of gestural expression, or dynamics [13,14,15], these sounds perhaps function primarily to express the feelings, states, and attitudes of the speaker [9,16] and then secondarily to support the text, or at least to function in parallel with it.

Being very frequent, and effectively 'transparent' with respect to the propositional content of the discourse, the prosodic features of these speech gestures can be easily detected and compared. In addition to obvious variation in duration and intonation they are also marked for 'tone-of-voice' i.e., phonatory voice-quality characteristics. Being so frequent, they can be compared 'like with like' as the speaker's and listener's affective and discoursal states and relationships change and progress throughout the discourse. As we have shown previously [10], the prosodic aspects of these non-verbal speech sounds share much in common across different cultures and languages, and they may represent a basic form of pre-linguistic human communication.

Figure 3 shows the corresponding part of the dialogue segment presented in Figure 2 which has been bracketed to highlight the frequently-repeated speech segments. Here we see the Japanese speaker's utterances and can combine them with those of her Chinese partner to reproduce the conversation segment. Some potentially ambiguous utterances can thereby be disambiguated by use of the textual content of the surrounding utterances, but a large number remain functionally indeterminate from the transcription alone. They are not at all ambiguous when listening to the speech, and carry a considerable amount of discourse information.

The text in Figure 3 has been annotated by a computer program to indicate which utterances are unique (and therefore presumably convey more propositional content) and to bracket those which are subject to frequent repetition and hence act as potential carriers of affect or discourse-control information. Two types of repetition have been bracketed: (a) whole phrases that occur more than a threshold number of times in the corpus, and (b) phrasal chunks that form part of a larger, possibly unique, utterance but which

are frequently repeated anyway. The current setting of the repeated-pattern recognition program, arbitrarily takes more than 99 repeats throughout the corpus as the minimum threshold for bracketing, and yields 74,324 untouched utterances, 72,942 marked as repeated phrases, and 49,136 utterances including repeated phrasal segments. These thresholds were determined by trial and error and are not intended to be more than examples.

Taking some of the frequent repetitions from one of the corpus speakers as an example, we notice different strategies of usage according to differences in partner. This speaker (JFA) makes considerable use of "ah", "ano", "hai", and "un", but not equally with all partners (see Table 2). For example, when speaking with foreigners, she uses "hai" frequently (はい=yes (with all of the ambiguity that it can have in an English conversation)), but significantly less so when speaking with Japanese partners. She uses "demo" (でも=but) much more frequently with Japanese partners, and "ah" much less when conversing (in Japanese) with the English-native-speaker partners.

Such differences may reflect interpersonal relationships, personal characteristics, or cultural peculiarities, but perhaps more interesting is the considerable variety of pronunciations within each utterance type, reflecting the speaker's interest, state-of-mind, and types of participation in the discourse.

3.2. Physical Characteristics of Repeated Segments

It is a central tenet of this paper that these repeated segments function to carry affect-related and interpersonal information in parallel to the linguistic content of the message. They do this by means of small but consistent variations in such acoustic characteristics as tone-of-voice, spectral tilt, pitch range and excursion, speaking rate, laryngeal and phonatory setting, etc. In this section we will examine some of these physical characteristics. By being so frequent and repetitive, the transparent speech gestures allow a listener (even one not yet familiar with the particular speaker's traits or habits) to make comparative judgements about the speaker's emotional and affective states and stances and to interpret subtle nuances in the speech by means of the prosodic cues hereby revealed.

Table 3 illustrates some differences in pitch range (i.e, the amount of variation in the f0 or fundamental frequency of the voice throughout the utterance) and voice energy (signal power in decibels) for three representative but randomly-selected sample speech gestures taken from speaker JFA's conversations with six different partners.

These data show that the speaker's basic acoustic settings and the amount of physical energy used in each utterance vary not just according to utterance type, as would be expected, but also according to the listener (and presumably according to the context of the conversations).

Table 2. Counts for some frequently-repeated simple utterances from one speaker to six partners. The table illustrates differences in usage strategies for these utterances.

JFA:	CFA	CMA	EFA	EMA	JFB	JMA
a,a–	143	145	88	89	138	170
ano	224	277	221	176	209	266
demo	41	24	31	17	89	134
e–	48	51	37	25	74	94
hai	2932	2234	2181	3239	72	33
un,un	1029	546	585	1190	909	1037

Table 3. F0 range (f0r) and mean (f0m) values in Hz and Power range (pwr) and mean (pwm) values in dB for three sample utterances (ah, umm, and ano) from speaker JFA according to differences in conversational partner

"a,a–"	CFA	CMA	EFA	EMA	JFB	JMA
f0r	125	181	266	232	234	241
f0m	201	214	220	192	206	198
pwr	28	29	29	28	31	31
pwm	38	39	36	35	42	41
"un,un"	CFA	CMA	EFA	EMA	JFB	JMA
f0r	154	152	182	181	161	141
f0m	172	175	162	145	198	174
pwr	28	29	27	26	29	27
pwm	37	40	36	35	42	39
"ano"	CFA	CMA	EFA	EMA	JFB	JMA
f0r	106	113	161	154	169	155
f0m	131	136	142	133	156	149
pwr	27	28	28	27	31	29
pwm	38	40	37	36	42	39

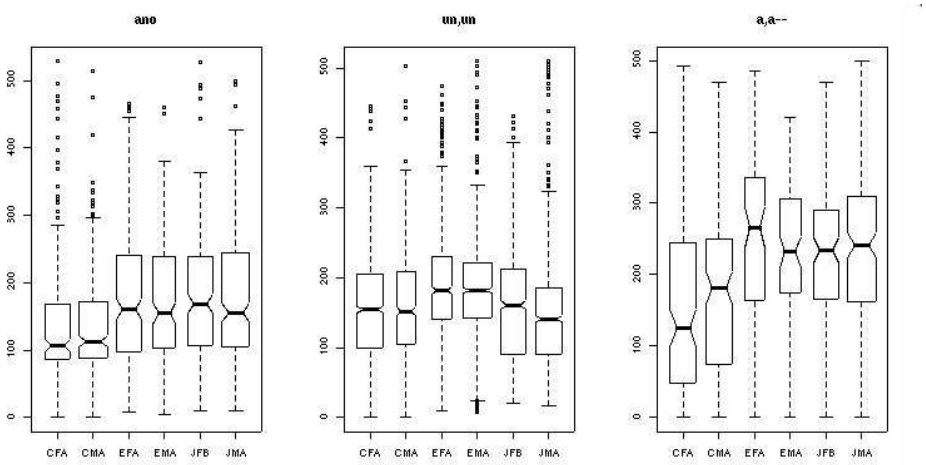

Figure 4. Plots of Pitch Range (amount of variation in the fundamental frequency of the voice) for three utterances from speaker JFA when conversing with six different partners. The width of the boxes is proportional to the number of tokens. Differences are significant at the 5% level if the notches do not overlap. The vertical axis shows pitch range in Hz.

Figure 4 takes a subset of this data (fundamental frequency contours for the utterance "un,un") and plots a representation of the 'shape' of each utterance by showing averaged f0 values for each progressive third of the utterance. Again we see that there is considerable variation, but that the variation between contours for different types of conversation partner is greater than that between utterances within a given set of conversations.

The data show that the speaker's basic acoustic settings and amount of physical en-

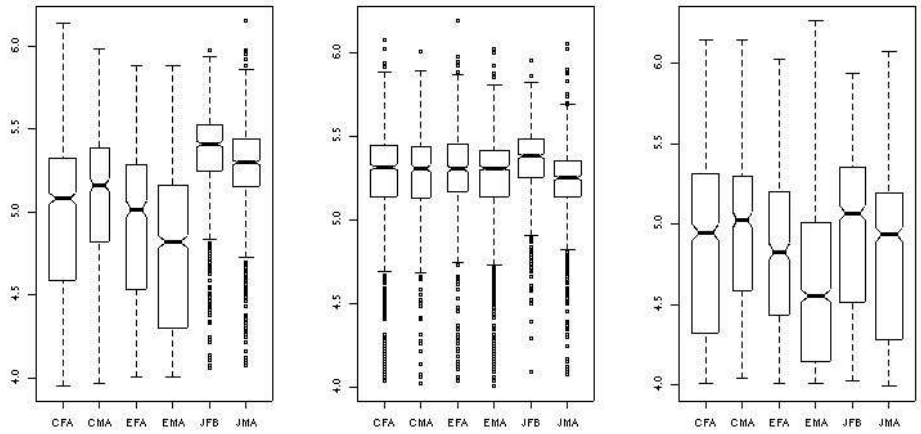

Figure 5. Fundamental frequency contours differ according to the listener. The left-hand plot shows average f0 values for the initial third of the utterance, the middle plot for the middle third, and the right-hand plot shows average f0 values for the final third of each utterance. Plots show'averaged contours' for all samples of the utterance "un,un" factored by partner. Japanese partners evoke a high initial contour, and English-native-speakers a lower fall at the end, though all countours appear to pass through the same high range of values mid-utterance.

ergy used in each utterance vary not just by utterance, as would be expected, but also according to the listener (and presumably according to the content of the conversations). Figure 5 takes a subset of this data (fundamental frequency contours for the utterance "un,un") and plots a representation of the 'shape' of each utterance by showing averaged f0 values for each progressive third of the utterance. Again we see that there is considerabe variation, but that the variation between contours for different types of conversation partner is greater than that between utterances within a given set of conversations.

It is apparent that Japanese partners evoke a high initial contour, and English-native-speakers a lower fall at the end, though all contours appear to pass through the same high range of values mid-utterance. The fact that these differences appear more related to partner than to local contextual differences implies that a higher-level of socially-inspired prosodic processing may be taking place; i.e., that a level of social interaction is influencing the prosodic contour just as the linguistic relations influence it a lower level.

4. Discussion

"In human communication a great deal of failure comes about not because information has been lost in transmission but because the sender is unable to express what he has to say, or because the receiver is unable to interpret the message in the way intended." [18]

In written communication, great care is usually taken so that the structure of the text should clearly and unambiguously portray the meaning intended by the author. In speech communication, on the other hand, the interaction is in real-time, two-way, and often constructed on the spur of the moment. Little time is available for a careful planning of

the structure of a spoken utterance, and the resulting 'text' is often broken up and spread out among several sequential utterance segments that are interspersed with discourse-control and interpersonal stance messages expressed non-verbally.

There is no guarantee that the speaker is optimally expressing her intended meaning, not that the listener is optimally comprehending the speech stream. Instead, a constant stream of feedback and feedback-elicitation is necessary so that the information transfer may be optimised. Failure of communication comes about when this secondary stream is ineffective.

Allwood's theory of *Communication as Action and Cooperation* [19] prescribes the communicative activities of a sender and a receiver and provides a framework for their interconnection. However, in current speech technology, only the primary stream (the linguistic or propositional content) is currently in focus for speech processing. The notion of 'communicative acts' is secondary to that of textual content.

This text-based form of information processing may be adequate for the analysis of broadcast news, where the speaker is transmitting to a plurality of listeners as a remote audience which has no interactive potential in real-time. However, future speech technology must incorporate *both channels* of information (verbal as well as non-verbal) if it is to process real-time interactive human speech communication efficiently.

The speech data from the ESP_C corpus of conversational dialogues confirm that there is considerable prosodic variation on what are seemingly very simple but also very frequent utterances. This variation may also serve to indicate the speaker's relationship with the listener since it seems to vary more between conversational partners than between different utterance types.

We can see from the data that the lowest level of discourse information can be processed in a speech signal for the automatic annotation of discourse progress and for producing an estimate of speaker participation status. The auditory landscape of a dialogue contains fluctuating surfaces of sound whose characteristics provide cues to the interpersonal relationships and discourse participation of the conversing partners.

This background provides an element of the discourse in which *how* something is said is more important than *what* was said, and where the prosody of the non-verbal speech components provides a dynamic expression to the simple 'umms' and 'ahhs' that are more normally considered as noise. By the interplay of such feedback comments and their elicitation, conversational speech takes on its characteristic forms of expression and the interactive transfer of knowledge is achieved.

5. Conclusion

This paper has described how the lowest level of discourse information can be processed in a speech signal for the automatic annotation of discourse progress and for producing an estimate of speaker participation status.

In a semi-formal round-table meeting situation there is typically only one main speaker at any given moment, but several participants may be speaking simultaneously, expressing agreement (or otherwise), chatting, translating, etc., in addition to the main speaker.

We are currently performing research into technology to process this audio landscape in order to detect the main speaker and to categorise the competing forms of speech in a given situation.

Acknowledgement

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 analysis was carried out 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

[1] Harrigan, J.A. & Rosenthal, R., "Nonverbal aspects of empathy and rapport in physician-patient interactions", pp. 36-73. In P.D. Blanck, R. Buck, & R. Rosenthal (Eds.). Nonverbal communication in the clinical context. University Park, PA: The Pennsylvania Univ. Press. 1986.

[2] The SCOPE 'robot's ears' project homepage: http://feast.atr.jp/non-verbal

[3] Campbell, N., "Conversational Speech Synthesis and the Need for Some Laughter", in IEEE Transactions on Audio, Speech, and Language Processing, Vol 14, No.4, July 2006.

[4] Kendon, A.,"Movement coordination in social interaction: Some examples described". Acta Psychologica, Amsterdam, 32(2): 101 125. 1970.

[5] Condon, W. S., "Communication: Rhythm and Structure. Rhythm in Psychological, Linguistic and Musical Processes", J. R. Evans and M. Clynes. Springfield, Illinois, Charles C Thomas Publisher: 55-78. 1986.

[6] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 27, no. 3, pp. 305 317, Mar. 2005.

[7] Zhang, D.,et al., "Multimodal group action clustering in meetings"., VSSN'04, 54-62, 2004.

[8] Campbell, N., "A Multi-media Database for Meetings Research", pp 77-82 in Proc Oriental COCOSDA, 2006, Jakarta, Indonesia.

[9] Campbell, W. N., "Listening between the lines; a study of paralinguistic information carried by tone-of-voice", pp 13-16 in Proc International Symposium on Tonal Aspects of Languages, TAL2004, Beijing, China, 2004.

[10] Campbell, N., & Erickson, D., "What do people hear? A study of the perception of non-verbal affective information in conversational speech", pp. 9-28 in Journal of the Phonetic Society of Japan, V7,N4, 2004.

[11] Campbell, N., "Non-Verbal Speech Processing for a Communicative Agent", Proc Eurospeech, pp. 769–772, Lisbon, 2005.

[12] The JST/CREST Expressive Speech Processing project, introductory web pages at: http://feast.his.atr.jp

[13] D. McNeill, "Gesture, Gaze, and Ground", in Proc 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms, Royal College of Physicians, Edinburgh, UK. July 2005.

[14] Condon, W. S., "Synchrony Demonstrated between Movements of the Neonate and Adult Speech", Child Development 45: 456-462. 1974.

[15] Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M. & Sedivy, J. E., "Integration of visual and linguistic information in spoken language comprehension". Science, 268, 1632-1634. l995.

[16] Campbell, N., "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.

[17] Campbell., N, & Suzuki, N., "Working with Very Sparse Data to Detect Speaker and Listener Participation in a Meetings Corpus", in Proc Language Resources & Evaluation Conference, Genova, Italy, 2006.

[18] Warner, T., "Communication Research", Vol. 19 No.1, p. 52-90 Communication Skills for Information Systems. London. Pitman Publishing, 1996.

[19] Allwood, J., "Linguistic Communication as Action and Cooperation", Goteburg Monographs in Linguistics, Goteborg University, Department of Linguistics, 1976.