ARCHIVES OF ACOUSTICS
32, 4, 803–814 (2007)

THE ROLE AND USE OF SPEECH GESTURES IN DISCOURSE

Nick CAMPBELL

(1) National Institute of Information and Communications Technology

(2) Spoken Language Communication Research Laboratory
Advanced Telecommunications Research Institute International
Keihanna Science City, Kyoto 619-0288, Japan
e-mail: nick@nict.go.jp, nick@atr.jp

(*received October 16, 2006; accepted February 26, 2007*)

This paper describes how a secondary level of discourse information can be processed in a speech signal for the automatic annotation of discourse progress and for producing an estimate of a speaker's 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. 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. The technology exists to recognise these discourse events, but we still lack a model of their function in the mutual transfer of information through speech interaction.

Keywords: speech recognition, audio landscape, speech gestures.

1. 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]. However, within the context of speech technology, this two-way characteristic of speech communication is not yet sufficiently considered. Speech synthesis produces audible text but usually in a broadcast-only mode wherein the presence or reaction of a listener is not considered. Speech recognition produces text from speech, but with no consideration of *how* it has been said, on the assumption that the meaning of an utterance is sufficiently defined by the sequence of words it contains. We thus have well-developed devices for the generation and recognition of propositional content in verbal speech, but very little is yet known about methods for processing the non-verbal content in a dialogue or for determining the intentions of a speaker when making a given utterance.

In written communication, great care is taken that the structure of the text should clearly and unambiguously portray the meaning intended by the author. In spoken communication, on the other hand, the interaction is real-time, two-way, and often spur-of-the-moment. Little time is available for the careful planning of the structure of a speech utterance, and the 'text' is often broken up and spread out among several sequential utterance segments that are interspersed with discourse-control and interpersonal stance messages expressed through non-verbal speech.

There is no guarantee that the speaker is optimally expressing his or her intended meaning, nor 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* [2] 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. The secondary stream of information communicated in a speech signal is that of speaker intention and attitude, which can be signalled through modulations in the prosody of the utterance. This secondary channel of information exchange through non-verbal speech utterances also enables feedback from the listener to help the speaker both to deliver content more efficiently, and at the same time to be reassured of the degrees of success in information transmission.

Future speech technology must incorporate *both streams* of information if it is to process real-time interactive human speech communication efficiently. This paper proposes that the secondary stream of speech information is carried most efficiently on small fragments of speech that are frequently repeated throughout a discourse, enabling otherwise imperceptible changes in vocal settings to be noticed through a 'comparison of like with like'. The paper describes some of our current work towards the recognition and modelling of speaker intentions underlying such repetitive non-verbal speech utterances in a dialogue context. The work is based on an analysis of a very large corpus of spoken interactions captured under extremely natural situations [3, 4].

The paper presents a view of speech interaction as not only facilitating the exchange of linguistic or propositional information, but also providing a mechanism for the display of affect and of interpersonal or social relationships. By incorporating such non-verbal speech content in our technology, we aim to provide for the two-way interactions between speaker and listener that are vital for facilitating the transfer of meaningful and situated discourse content.

2. The audio landscape

The 'audio-landscape' of a discourse enables a participant or observer to estimate the types of interaction and the relative status of participants without the need for a

content-based analysis of any particular utterance or sequence of utterances. By just watching what is happening in a discourse, 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), 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 and CONDON [5, 6] and the more recent 'meetings' findings [7–9]. In other words, even a foreigner who has no understanding of what is being said can often make an intelligent guess about the functional states (i.e., what is going on) in a dialogue, at the interpersonal level.

Furthermore, if we have access to the sound then we can make an intelligent guess about how the listeners are reacting to the content of the speech, even if (like the foreigner) we do not understand the content of the speech itself. Laughs, nods, grunts, etc., indicate the degree to which the listener is attentive, synchronised with the content of

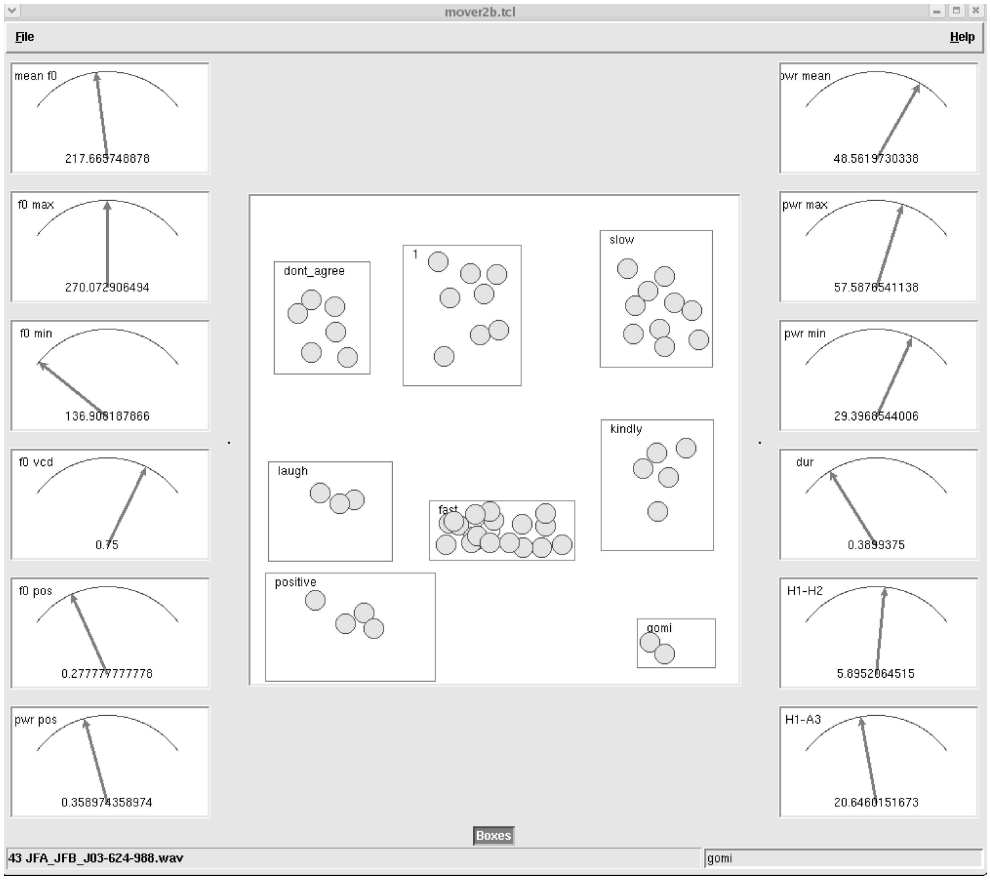

Fig. 1. The Mover software that enables listeners to arrange speech samples according to their perceived similarity and then to label the clusters accordingly. Here the speech data are laughs. The acoustic meters shown on the right and left are not usually displayed when labelling is in progress, but can be switched on by a supervisor to visually check correlation of acoustic features within groups.

the discourse, and in relative states of agreement with it. This much can be determined from the non-verbal content.

Figure 1 shows a screenshot of the Mover software [10] that we used to confirm this claim and to provide subjective labels for each class of utterance [11]. Subjects are initially presented with speech samples represented by a row of dots aligned in a row from top-left to bottom-right along the main diagonal of the screen. Their task is to move these dots into other areas of the screen, so that like is placed close to like, and then to identify these areas with a descriptive name of their own choice. The process an open-ended one, is much like sorting books or gramophone records and can be very subjective in terms of the categories selected. In [11], we presented Japanese speech samples to Korean, French, and North American listeners who had no previous experience of Japanese, and found considerable agreement in their reactions and in the labels they chose to describe the supposed intentions of the speakers of the utterances.

2.1. Discourse participation

We are currently attempting to describe this 'audio landscape' in order to detect the main speaker in a given discourse situation (especially in a semi-formal 'meetings' environment [12, 13]) and to categorise the competing forms of speech as the interaction progresses. 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.

In a semi-formal round-table meeting situation we have found that there is typically only one main speaker at any given moment [14], but that several participants may be speaking simultaneously, expressing (dis-)agreement, chatting, translating, etc., in addition to the main speaker. In a two-person dialogue, the amount of noise is reduced, but the types of interaction can be considered equivalent if we also take into account the attentional states of the conversational partners. The data reported below are from telephone conversation, rather than meetings, but we assume that the findings from the one can gereralise to the other.

Speakers laugh and use other such non-verbal gestures frequently in natural conversations. These data provide much of the audio landscape of a conversation.

2.2. Speech gestures

Figure 2 shows part of a dialogue segment which has been bracketed by a computer programme to indicate the frequently-repeated speech segments. The content of the dialogue is not relevant here, and we focus instead on its form. Column 7 in the figure indicates graphically the class of utterance, using "– –>" to show propositional content, "= =>" to show frequently repeated non-verbal utterances (indicated in the text by <[*square*]> brackets) and "+ –>" to indicate a mixed-type utterance that contains a fragment that is very common in the speech of that speaker (indicated in the text by double <<*angle*>> (which can be recursively assigned)). These bracketings are de-

termined automatically, based on the frequency of each fragment in the history of the discourse.

```
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 ]>
```

Fig. 2. Part of a dialogue, showing frequent utterances ($n \geq 100$) in <[*square*]> brackets, and frequent segments ($N \geq 100$) as part of longer utterances in <<*angle*>> brackets, which may be embedded. Columns show (from left to right) Speaker, Listener, Conversation number, Start time, Duration (in seconds) and Delay (milliseconds) from end of previous utterance. A symbol indicating class of utterance (see text), and the text of the utterance in Japanese. JFA is Japanese female speaker from Group A, CFA is a Chinese female speaker from the same group. C01 indicates their first conversation. The text is not translated because it is the repetitions that are of more interest here, rather than the linguistic content. The paper argues that these very frequently repeated fragments (bracketed in the text) carry a separate stream of information through their prosody.

Certain types of utterances, such as would be translated as “yeah” or “uhuh” in English are extremely common and indicate listener attention. If we exclude these feedback utterances (i.e., just listen to those utterances marked with “– –>” in Fig. 1), 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 pri-

marily non-verbal utterances (marked with "= =>" 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 of these fragments to make an inference about the function of each utterance in the discourse without knowing its specific lexical meaning.

Table 1 lists the 100 most-frequently bracketed fragments from a corpus of 200,000 dialogue utterances recorded from six people's telephone conversations where there was no face-to-face contact or visual information. We can see (even those of us who cannot yet read Japanese) that repetition is a common identifying characteristic of these frequently-repeated utterances. The same syllable (shown here by a Japanese character or character sequence) in more than half of the cases. If we expand this list to include

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 moraic time slot indicated here 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. They would translate into English as "umm", "er", "huh", "haha", etc., Note that these few samples alone account for more than a third ($n = 72,685$) of the 200,000 utterances in the corpus. Fewer then half ($n = 92,541$) of the utterances were unique according to this reckoning.

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

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 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), one simple repetition ("umm": うん, うんうん, うんうんうん), one complex repetition ("ha!": ハ, ハハ, ハハハ, ハハハハ, ハハハハハ, 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 (and typically longer) variants, then we find that more than half of the utterances in the corpus are of this non-verbal type (not its usually found in a standard dictionary).

Being very frequent, and effectively 'transparent' with respect to the propositional features of the discourse, the prosodic, 'tone-of-voice', and phonatory voice-quality characteristics can be measured and compared, like with like, as the speaker and listener states change and progress throughout the discourse. The highly frequent nature of their repetitions makes them ideal for use as a carrier of information related to affective states and discourse roles. As we have shown in [11], the prosodic aspects of these non-verbal speech sounds share much in common across different cultures and languages, and may represent a basic form of pre-linguistic human communication.

3. Features of non-verbal speech

Unlike regular lexical items which have a fixed form and a variable prosody depending on their contexts, these non-verbal 'speech gestures' rather seem to have a fixed prosodic expression (or prosodic dynamic) and a variable form, expanding freely to meet the requirements of the prosodic dynamics that they serve to carry. Like bodily gestures, which have only a few finite forms but considerable freedom in gestural expression (or gestural dynamics), these sounds perhaps function primarily to express the feelings, states, and attitudes of the speaker [15] and then only secondarily to support the propositional content.

To track the changes over time of these short and frequent utterances throughout a discourse, we use a combination of acoustic features as described below. We have found from experience that seven principal components derived from fourteen acoustic features are adequate to model the prosodic and acoustic characteristics of these utterances, and that the first three components account for almost half of their variance (see Table 2).

The acoustic meters shown on the left and right of the main screen in Fig. 1 indicate the settings of each speech sample for acoustic variables representing the pitch, power, spectrum, and duration of each utterance. The meters show the mean, maximum, minimum of power (rms amplitude) and pitch (f0), the positions of the f0 peak and the maximum energy in each utterance, measured as a percent-age distance from 0 (beginning) to 100 (end of utterance), the amount of voicing throughout the utterance, the energy values at the first and second harmonics, the third formant, and the spectral tilt

Table 2. Results of the Principal Component Analysis. Here, SD stands for Standard Deviation, PV for Proportion of Variance, and CP Cumulative Proportion of the variance that has been accounted for.

Importance of the components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
SD	1.65	1.53	1.38	1.32	1.12	0.96	0.89
PV	0.19	0.16	0.13	0.12	0.08	0.06	0.05
CP	0.19	0.36	0.49	0.62	0.71	0.78	0.83

	PC8	PC9	PC10	PC11	PC12	PC13	PC14
SD	0.83	0.74	0.71	0.61	0.29	0.23	0.0004
PV	0.04	0.03	0.03	0.026	0.006	0.004	0.0001
CP	0.88	0.92	0.96	0.98	0.99	1.00	1.00

(after HANSEN [16]), as well as a measure of speaking rate or normalised duration of the utterance. These measures are averaged across the whole of each utterance giving only a general indication of the prosodic characteristics of 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. These are indicated by the abbreviations “fmean”, “fmax”, “fmin”, “fpct”, “fvcd”, “pmean”, “pmax”, “pmin”, “ppct”, “h1h2”, “h1a3”, “h1”, “a3”, “dn” in Tables 2 and 3.

Table 3. Showing the precise relationship between each 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

Table 2 provides details of the principal component analysis, showing how much of the variation is covered by each component and Table 3 shows how the individual acoustic measures are mapped by the components in the pca reduction. We can see that approximately half of the variance is covered by the first three components, and that more than 80% is accounted for by the first seven. By tracking the changes in these variables, we can follow the progress of speaker involvement in the discourse.

From Table 3, we can see that the first principal component relates closely to fundamental frequency, the second to spectral tilt, and the third to energy in the speech signal. It is coincidental that all three components appear positive when the speech is warm and friendly, and negative when the speech is tense or stressed, but the changes in these values can be used to track speaker involvement throughout the discourse. Although values will be different for each individual phonetic content and type of utterance, they can be compared easily when the same utterance fragment is repeated frequently at intervals throughout the conversation.

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" [17].

It is essential that the technology to process speech be able to also process the signals that cue the intended interpretation of the message content. It is also important that the technology be able to produce the appropriate backchannel sounds so that a human-machine interaction may progress smoothly. We can hope that a future speech technology may be able to function interactively with a speaker to jointly contribute to the construction of the message.

Our current work uses a small number of acoustic features and tracks the way they change over time by comparing very frequent non-verbal utterances, or speech fragments, that play only a secondary role in the conversation, as a first step in a longer process to provide information about the intentional, attentional, and affective states of the speaker. We have shown that human listeners are capable of interpreting these speech sounds to arrive at an understanding which appears to be universal. There are of course local culture-dependent and language-dependent differences in how the individual sounds are used, but the basic prosodic elements of voice quality, pitch excursion and speaking rate are human characteristics which appear to be language-independent when used in this secondary way.

We assume that since these non-verbal utterances are characterised primarily by their prosodic contours, by tracking their development over time we will be able to relate these changes to the features that human listeners perceive. We already have statistical mapppings between the acoustic features and the perceived labels, but determining the precise relation between the acoustic changes and the auditory perceptions remains as ongoing work.

Concurrent work [18, 19] using the same model is being carried out for 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

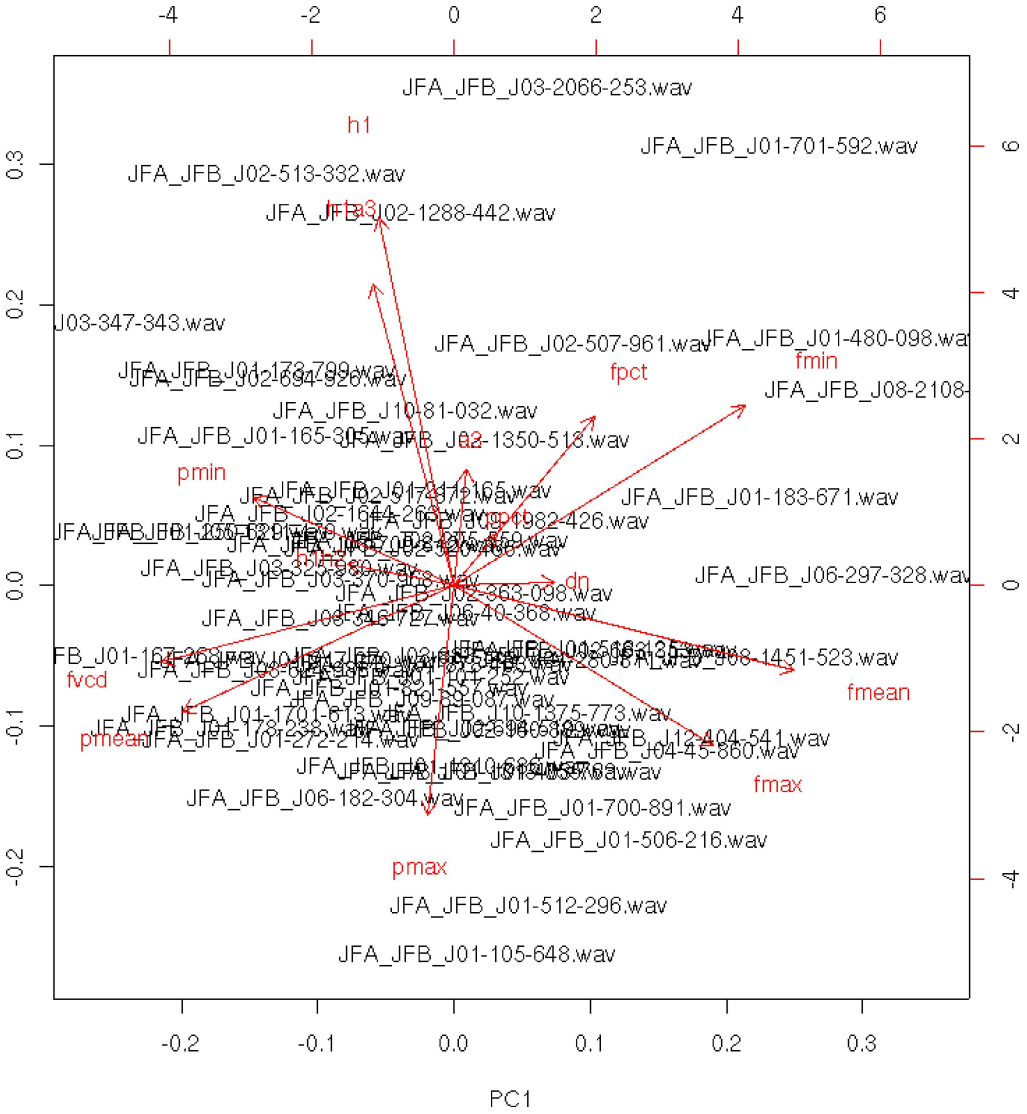

Fig. 3. Results of a Principal Component Analysis showing how the acoustic features cluster, and how individual speech waveforms relate to the various acoustic parameters. Only the first 2 principal components are plotted in this figure. Axes at the top and right show normalised scales for the acoustic features, while those on the left and bottom show values for the individual speech waveforms in the first and second component spaces.

speech sounds 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 (i.e., listener) states in a discourse. By using the pca scores as a feature in the unit selection, appropriate variants of these non-verbal sounds can be selected.

5. Conclusion

The auditory landscape of a dialogue contains fluctuating surfaces of sound whose characteristics provide cues to the social relationships, affective states, and discourse roles 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 by their elicitation, conversational speech takes on its characteristic forms of expression and the interactive transfer of knowledge is achieved.

In addition to the propositional content, in parallel, there is a level of interpersonal communication that takes place through the use of non-verbal speech sounds and that is signalled primarily by variations in their prosody. This paper has described how this secondary level of information can be processed in a speech signal for the automatic annotation of discourse progress and for producing an estimate of speaker participation status.

Acknowledgment

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] KEENAN J., *Intellectual Capital and Knowledge Management in Organization*, Doctoral course at MidSweden University, April 1999.

[2] ALLWOOD J., *Linguistic Communication as Action and Cooperation*, Goteburg Monographs in Linguistics, Goteborg University, Department of Linguistics, 1976.

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

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

[5] KENDON A., *Movement coordination in social interaction: Some examples described*, Acta Psychologica, **32**, 2, 101–125 (1970).

[6] CONDON W. S., *Communication: Rhythm and Structure. Rhythm in Psychological, Linguistic and Musical Processes*, pp. 55–78, J. R. Evans, M. Clynes [Eds.], Charles C Thomas Publisher, Springfield, Illinois 1986.

[7] MCCOWAN I., GATICA-PEREZ D., BENGIO S., LATHOUD G., BARNARD M., ZHANG D., *Automatic analysis of multimodal group actions in meetings*, IEEE Trans. on Pattern Analysis and Machine Intelligence, **27**, 3, 305–317 (2005).

[8] ZHANG D., *et al*., *Multimodal group action clustering in meetings*, VSSN'04, 54–62, 2004.

[9] KATOH M., *et al*., *State estimation of meetings by information fusion using bayesian network*, INTERSPEECH2005, 113–116, 2005.

[10] Mover – a software programme written in Tcl/Tk for the categorisation of speech utterances. Made available by ATR at http://feast.atr.jp/software/, 2005.

[11] CAMPBELL N., ERICKSON D., *What do people hear? A study of the perception of non-verbal affective information in conversational speech*, Journal of the Phonetic Society of Japan, **7**, 4, 9–28 (2004).

[12] CAMPBELL W. N., *A Multi-media Database for Meetings Research*, [in:] Proc. Oriental COCOSDA, pp. 77–82, Jakarta, Indonesia 2006.

[13] CAMPBELL W. N., *Non-Verbal Speech Processing for a Communicative Agent*, Proc. Eurospeech, pp. 769–772, Lisbon 2005.

[14] 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.

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

[16] HANSON H. M., *Glottal characteristics of female speakers: acoustic correlates*, J. Acoust. Soc. Am, **101**, 466–481, 1997.

[17] WARNER T., *Communication Research*, Communication Skills for Information Systems, Vol. 19, No. 1, p. 52–90, Pitman Publishing, London 1996.

[18] CAMPBELL N., *Conversational Speech Synthesis and the Need for Some Laughter*, IEEE Transactions on Audio, Speech, and Language Processing, **14**, 4, July 2006.

[19] CAMPBELL N., *Speech Synthesis and Discourse Information*, [in:] Proc. Fifth Slovenian and First International Language Technologies Conference, October 9-10, Ljubljana, Slovenia 2006.