Non-Verbal Speech Processing for a Communicative Agent

Nick Campbell

ATR Media Information Science Labs,
Kyoto 619-0288, Japan
nick@atr.jp

Abstract

A believable life-like agent must appear to understand what is said to it, or what is being said around it, even if this is not actually the case. It must be able to follow a conversation and to understand what is happening in a discourse, even though the verbal content of the dialogue may be too complex (or too noisy) to be recognised. This paper describes the platform design and details the description language being used for the 'SCOPE' non-verbal speech-processing research currently being carried out at ATR for such a communicative robotic agent.

1. Introduction

A traveller in a foreign country may not have a command of the local language but that does not necessitate a complete block of communication; indeed, quite the opposite, in a wide range of communicative situations a person can both understand and be understood even without the use of linguistic communication [1, 2]. Much of the information in this type of extra-linguistic discourse is gestural, with hands and face being particularly eloquent, but a large part is also conveyed by tone-of-voice and choice of speaking-style. In natural human spoken interactions, the listener makes use of a broad range of verbal and non-verbal information sources in addition to that of the linguistic channel, and this paper describes work to simulate this processing of affective and non-verbal discourse information for intelligent robotic systems.

Previous work [3] focussed on the analysis of a very large corpus (in excess of 1000 hours) of natural everyday conversational speech recorded over a period of 5 years from a small number of volunteer subjects during their normal daily interactions. The recordings were made using high-quality head-mounted microphones worn throughout the day by the subjects as they went about their daily affairs [4]. The resulting speech was digitised, transcribed, and annotated for affective and discoursal features. One major result of that work was the finding that more than half of the transcribed utterances were primarily non-verbal in content [5]. By analysing the types of non-linguistic speech sounds that were so common in the ESP corpus, we found that they functioned to convey a parallel channel of affective information that served to complement the linguistic information and to maintain discourse flow while at the same time establishing speaker and listener relationships [6].

This paper describes work being carried out in a follow-up project [7] to study the extent to which the flow of discourse information can be processed from non-verbal (speech and gesture) information. We have limited our research context to that of a six-member meeting, and track the flow of discourse and participant relationships during the meeting in order to (a) produce a listing of those parts of the meeting for which a transcription would be necessary, and (b) to produce a flow analysis independently of any linguistic information. We envisage this technology being used in social conversational agents, robots, ubiquitous computing environments, and ambient intelligent spaces.

2. Non-verbal Speech Processing

The 'Robot's Ears' project (as it is informally known) is part of the Strategic Information and COmmunications R&D Promotion ProgrammE (SCOPE) initiative to fund research into next generation human interface & contents technology. While focussing on the recognition and processing of non-verbal information related to speech, this project differs from the ESP work described above in that it adds a visual component, analysing body movements as well as the audio stream. It utilises technology generated during the JST/CREST ESP research and extends our understanding of non-verbal speech processes to include visual information.

2.1. A Platform for Dialogue Processing

A small 360-degree camera is placed in the centre of the meeting table and is surrounded by a ring of directional microphones, pointing outwards, to collect a stream of audio-visual information from which to characterise the discourse events of the meeting. The video signal is of relatively low resolution (see figure 1), so fine details such as eye-gaze and direction are not available to the system in its present design (and will perhaps not be necessary). Instead, gross movements are detected from the skin tones and a set of primitive features describing the body, hand and head movements is produced automatically.

The output from a co-axial ring array of microphones (figure 2) is synchronised by use of a multi-input digital-to-analogue firewire device to produce a multichannel signal which is subsequently decomposed into the individual audio tracks. The amplitude (rms power) of each waveform is calculated using a sliding 1024-point Hamming window with a 50-millisecond step-size and the relative amplitude of each line is used to provide an indication of the local variations in overall sound quality around the table. From this information, in conjunction with the video primitives, we are attempting to produce an estimate of each speaker's activity throughout the meeting.

Currently several hardware configurations are being tested, with a close-to-perfect audio signal being captured by a set of four Sennheiser MKH-60 P48 shotgun microphones arranged in a central windmill configuration to provide the reference signal, in conjunction with six small Audio-Technica radio-microphones, for intermediate quality sound capture, not worn on the body but mounted on the tables in front of the participants, beside an array of domestic Sony ECM-Z590 stereo microphones which are closer to the consumer-level recording quality that we envisage using for the final product.

Figure 1: A 360-degree camera provides a view of all participants at the meeting, and although the resolution is low, their body movements can be tracked using skin-tone-sensitive image processing algorithms (see fig 3)

We are testing 2 miniature professional cameras in addition to a high-end domestic quality cam-corder (Sony DCR-HC1000) for video capture. The cam-corder would allow us ultimately to provide a cheaper version of the technology, but we have not yet succeeded in realising good image-detection from this source, which is currently regarded as a back-up device to facilitate record-keeping and subsequent human labelling of the video data, having its own integral sound with 2 or 4-track recording.

Output from the video capture device is set to 15 frames per second, for tracking face and hand movements only, and the audio is summarised at 1 frame per second as explained in the following section. A smoothing algorithm is used to keep track of the video object IDs; if an object is matched from one frame to the next one then the previous ID is kept, otherwise a new ID is given. However, this algorithm will need further improvement as, at present, object IDs are kept coherent only for a maximum of around 400 frames (30 seconds). The centre of gravity of each object is provided in a low bit-rate output data stream, along with information describing the object motion in rectified coordinates. Positive X motion indicates that the object moves to the left (e.g. the person looks or moves a hand to the left), positive Y motion indicates that the object moves up (e.g. the person moves his or her head up), and vice versa (figure 3).

2.2. Multimodal Pattern Recognition

The output from the manual labelling of the audio stream is shown in figure 4. The six columns represent the speech activities of the six participants at the meeting, and the rows present changes in the data sampled once every second. Times at which no event is labelled have been omitted for reasons of space (e.g., rows 3320-3322).

Six categories of speech event are annotated, on the basis of human judgement. A vertical bar (e.g., speaker B 3319-3340, speaker D 3324-3342) indicates a continuous stream of speech, such as might require transcription or translation, ‘y’ and ‘n’ represent ‘yes’ and ‘no’ respectively, and ‘p’ and ‘t’ represent private conversations, ‘t’ being reserved for ‘translation’ or explanation. Laughter is noted using the letter ‘w’.

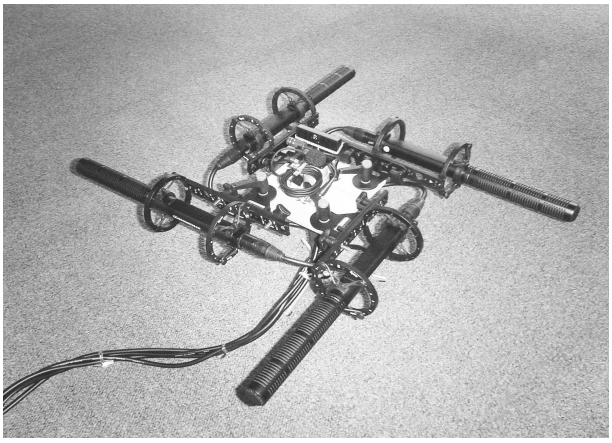

Figure 2: The microphone windmill— by placing direction-sensitive microphones at the cardinal positions, we are attempting to track the location of each sound source

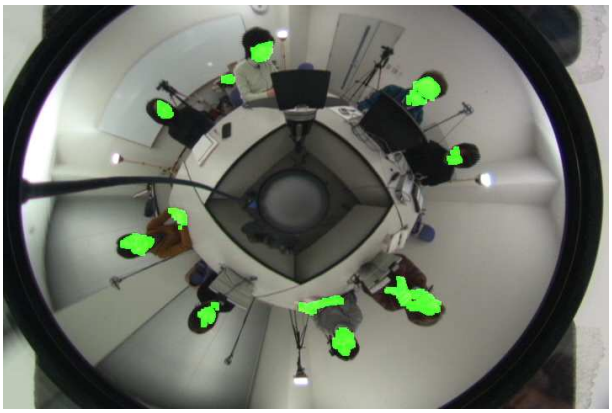

Figure 3: After image processing, we obtain a map of the areas showing skin-tones. Here we see moving parts of both Japanese and European participants being tracked

It is relatively easy for a human observer to follow the flow of the discourse from this simple representation, although it is not yet possible to obtain such fine granularity of information directly from the audio signals (this is ongoing work that will be presented separately). Our present goal is to derive the finer classification of the long and short speech utterances by use of video information taking into account the movements of both head and hands for each participant.

The audio stream provides us with information about who is speaking, and about the duration of each utterance. A stream of continuous speech is assumed to be either public or personal, with personal talk further sub-classified into side-chat (private, p) or translation (t) which is directly related to the main flow of public speech, such as explanation in another language. Short utterances are assumed to be backchannels or laughter. Backchannels are further sub-classified into either positive or negative variants, assuming that a ‘neutral’ backchannel can be taken as positive, since it encourages continuation of the discourse.

	participants					
time	A	B	C	D	E	F
....	.	.	.	.	.	.
3315	–	y	–	–	–	–
3316	–	–	–	–	–	–
3317	–	y	|	–	–	–
3318	–	y	|	–	–	–
3319	–	|	–	–	–	–
3323	–	|	–	–	–	y
3324	–	|	–	|	–	y
3325	–	|	–	|	–	–
3327	–	|	–	|	–	y
3328	–	|	–	–	–	y
3332	|	|	–	|	–	–
3333	–	|	–	|	–	–
3335	–	|	–	|	–	–
3336	|	|	–	|	–	–
3340	–	|	–	|	t	–
3341	–	w	–	|	t	–
3342	–	–	–	|	t	–
3345	–	–	–	y	t	y
3346	–	–	–	y	t	–
3347	|	–	–	y	t	–
3348	|	–	–	y	t	–
3350	|	–	–	w	t	–
3351	|	w	p	w	t	w
3352	–	–	p	w	t	–
3353	–	–	p	w	t	–
3355	–	–	–	w	t	–
3356	–	–	–	|	t	–
3357	–	–	–	–	t	–
3360	–	y	p	–	t	–
3362	y	y	p	–	t	–
3363	–	–	p	–	t	–
3365	y	–	p	–	t	–
3366	–	–	p	–	t	–
3367	y	–	p	–	t	–
3369	–	–	–	–	t	–
3371	–	–	–	–	–	–
3372	–	w	p	–	–	–
....	.	.	.	.	.	.
sex	m	f	m	f	f	m
seniority	s	j	j	m	s	m

Figure 4: labels - showing approximately one minute of dialogue between six participants (A-F), three male (m) and three female (f), two senior (s), two junior (j), and two mid-level (m) in status. The numbers indicate time in seconds into the meeting. See text for key to the labels

We can see from the figure that public utterances tend to be proceeded by short utterances (laughing or agreement) and close similarly, with either a laugh or an affirmation. Reading across the columns in figure 4, we can see the general groupings. For example, there is high activity at time 3351, with every participant speaking at once, and three laughing, at the end of speaker A's activity stream. This is followed by a lull, up to time 3357, when only D is left speaking, followed by a renewed burst of undirected speaking between times 3360-3367.

We can follow the flow of the dialogue as speaker B initiates a theme from time 3317, starting by agreeing to something from speaker C, and ending with a laugh at time 3341, then speaker D takes up the theme, speaking in parallel from time 3324. Speaker E starts explaining to speaker F (a visitor who speaks very little Japanese) and continues beyond time 3351 (a point of general agreement?) during the private conversation between speaker C and (perhaps?) speakers B and A. The discourse action is easy to follow without listening to the words.

While no note has been taken of linguistic content in this analysis, we might want to listen to the speech of speaker E between times 3340-3369 for a summary, or that of speaker D between times 3324-3345 for an understanding (or a gist) of the content of this section of the meeting. This method of representing the dialogue flow of a meeting greatly reduces the amount of recordings for which a transcription is required, and at the same time gives an impression of the flow of the meeting, ultimately perhaps allowing us to estimate who was listening to whom at different times, and whether there was general agreement or interest at each dialogue event.

2.3. Incorporating Video Information

The labels shown in figure 4 were produced by human labellers (using Wavesurfer in video plugin mode [**?**]), listening to the various participants recordings separately, and the resulting label streams were aligned and merged to produce the display above. This illustrates the quality and types of information that we can hope to work with, and our main task at the present moment concerns the automation of this labelling process by statistical mapping between the raw and labelled data.

Whereas the audio streams provide information about who is speaking at each moment and for how long, they do not provide the finer details that are annotated here, such as 'main theme', 'private talk', 'backchannel', etc. This information must be generated from other cues present in the data.

Period of talk is the first level of simple information that can be made use of; allowing us to discriminate between short bursts and sustained periods of speech from each participant, and to make use of any synchrony of timing between and among the bursts. Laughter is a distinctive short burst and can be recognised or distinguished to a large extent from the audio and synchrony characteristics [8], but the difference between e.g., agreement and disagreement is more subtle. Here we need to include information from head and hand movements to help us disambiguate between the two.

The manually-obtained feature labels for the video stream are shown in figure 5 for the equivalent time period as the audio data above. We can see that speaker B's hands are together (b:both) at the start (while listening?), but that they move leftwards as she starts talking and then are relatively inactive while she speaks. Speaker A's head is down while he is listening, but moves left for his first contribution, and again right for his second. His head is down for a period at time 3356, but moves to the left at 3372 when speakers B and C make a contribution. We might infer from this timing that he is looking at them while listening.

Our present task is to learn the significant correlations between such movements that can be relatively simply detected by the audio- and image-processing algorithms and the events in the dialogue as they are annotated by the labellers. However, rather than attempt to map single movements to events in the discourse, which might be a rather simplistic approach, we are looking for synchrony of movements, so that when several participants move in unison, the direction and nature of these movements might lead us to detect their relationship to the sounds present at the time. This synchrony need not relate to events immediately present, but to events soon to occur, since pointing and looking, like clearing of the throat, are good indicators of forthcoming speech activity.

In order to take advantage of a god's-eye-view of the discourse, we do not attempt real-time processing at present, but process the label sequences and the audio-video data streams

with a wider window of several seconds in both directions from the current point of attention. We are testing both Support-Vector-Machines and Hidden Markov modelling of these data sequences but the results of this training will be presented separately, as the focus of this paper is more on the general approach taken to the problem and on the architecture of the development platform.

3. Summary and Discussion

The goal of this project is to determine just how much of a dialogue can be understood without resorting to linguistic information; i.e., not to understand the text of the meeting, such as would be provided by a transcription of its content, but to form a sense of its 'flow' and of the roles of the participants at various stages of the meeting.

The use of simple primitives such as direction of speech (who is talking), duration of speech (pragmatic function), synchrony of hand and head movements among the participants (who is listening) and direction of movement (who is grouping with whom) can function as a rich source of information from which to infer significant details about the flow of a dialogue and about the roles of the individual participants, and from which an emergent understanding of the discourse can be obtained. Of course, for a full understanding of the content, some speech recognition and text processing will eventually be necessary, but that is not our goal. We will be satisfied when we can point out the small subset of the whole discourse that is sufficient to summarise the main points.

Although it is still too early in the project to present an evaluation of the relative success of the various processing algorithms, this paper has focussed instead on presenting the methodology and describing the platform for the research. it is novel in that it makes use only of non-verbal information and that no recourse is taken to linguistic knowledge. We believe that it is currently very difficult to perform adequate speech recognition in a real-life conversational environment and maintain that it will continue to remain so for several years to come. However, even without knowledge of the content of the speech, we can still understand much about what is going on in the discourse, as the example data above has illustrated.

4. Conclusion

This paper has described ongoing work to process the flow of dialogue information for a conversational agent. The technology requires a 360-degree camera and an array of microphones, the video and audio signals from which are processed to produce an array of multimodal information that is mapped to higher-level primitives which are in turn mapped to discourse-state and flow tags. The output of the system is a list of time-related events describing the states and relationships of the participants with respect to the dialogue.

5. Acknowledgements

This work is partly supported by the Japan Science & Technology Corporation (JST), partly by the National Institute of Information and Communications Technology (NiCT), and partly by the Ministry of Public Management, Home Affairs, Posts and Telecommunications, Japan.

time	head						body						hands					
who:	A	B	C	D	E	F	A	B	C	D	E	F	A	B	C	D	E	F
....	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3314	-	r	l	r	l	-	b	f	-	f	-	-	b	b	b	r	l	b
3316	-	r	-	r	l	-	b	f	-	r	-	-	b	b	b	-	l	b
3317	d	r	-	r	l	r	f	f	-	r	-	-	b	b	b	-	-	b
3318	d	l	-	r	l	-	f	r	-	r	-	-	l	l	b	-	-	b
3319	d	l	-	r	l	r	f	r	-	r	-	-	l	l	b	-	-	b
3320	d	r	-	r	l	r	f	r	-	r	-	-	l	-	-	-	b	b
3322	d	r	-	r	l	r	f	r	-	r	-	-	l	-	-	-	b	b
3326	-	r	l	r	l	u	r	r	-	r	-	-	l	l	-	-	b	b
3327	l	r	-	r	l	u	r	r	-	r	-	-	l	r	-	r	b	b
3334	l	r	l	r	l	-	r	r	-	r	-	-	r	-	b	r	b	b
3337	-	l	l	r	-	-	-	r	-	r	-	-	-	-	-	r	-	b
3338	r	r	l	-	-	-	-	r	-	r	-	-	-	-	-	r	-	b
3339	r	r	-	-	-	l	r	r	-	r	-	-	-	-	-	r	-	b
3340	r	r	r	-	-	l	r	r	-	r	-	-	-	-	-	r	-	r
3341	r	r	r	-	r	l	r	r	-	r	-	-	-	-	-	r	b	r
3343	d	l	-	r	r	-	-	r	-	r	-	-	-	l	b	r	b	r
3344	d	r	-	r	r	-	-	r	-	r	-	-	-	b	b	r	b	r
3345	l	r	-	d	r	-	-	r	-	r	-	-	-	b	b	r	b	r
3354	l	r	l	d	-	l	-	r	-	-	-	-	-	b	-	l	-	r
3355	l	r	l	r	l	l	-	r	-	r	-	-	-	b	-	l	-	r
3356	d	r	r	r	l	l	-	r	-	b	-	-	b	l	b	l	-	r
3359	d	-	r	r	l	l	l	r	-	f	-	-	l	l	b	l	-	l
3360	d	-	r	r	l	l	l	r	-	f	-	-	l	b	b	l	-	l
3361	d	-	r	r	r	l	l	r	-	f	-	-	l	b	b	l	-	l
3372	l	-	r	r	r	l	l	r	-	f	-	-	l	b	b	l	-	l
....	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Figure 5: A section of the movement label data for the period shown in figure 4 - approximately one minute of dialogue between six participants (A-F), the letters u,d,r,l, indicate up,down,right,left; b:'back' for the body, but 'both' for the hands. The numbers indicate time in seconds into the meeting

6. References

[1] Campbell, W. 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, Vol 7, Num 4, 2004.

[2] Campbell, W. N., "Getting to the heart of the matter" (LREC Keynote Speech), Proc International Conference on Language Resources and Evaluation, V-vii-x, Lisbon, 2004

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

[4] Campbell, W. N., "Speech & expression; the value of a longitudinal corpus", pp.183-186 in Proc Language Resources and Evaluation Conference (LREC-04), Lisbon, Portugal, 2004.

[5] 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.

[6] Campbell, W. N., "Expressive Speech - Simultaneous indication of information and affect", pp.49-58 in *From Traditional Phonology to Modern Speech Processing"(Festschrift for Professor Wu Zongji's 95th birthday*), eds G.Fant, H.Fujisaki, J.Cao & Y.Xu, 2004

[7] http://feast.his.atr.jp/non-verbal

[8] Ohara, R., "Classifying Four Types of Laughter", Masters Thesis, Department of Applied Linguistics, NAIST, 2005.