



# Analysis of head motions and speech in spoken dialogue

Carlos T. Ishi<sup>1</sup>, Hiroshi Ishiguro<sup>1,2</sup>, Norihiro Hagita<sup>1</sup>

<sup>1</sup> Intelligent Robotics and Communication Labs., ATR, Kyoto, Japan

<sup>2</sup> Department of Adaptive Machine Systems, Osaka University, Japan

carlos@atr.jp, ishiguro@ams.eng.osaka-u.ac.jp, hagita@atr.jp

## Abstract

With the aim of automatically generating head motions from speech, analyses are conducted for verifying the relations between head motions and linguistic and paralinguistic information carried by speech. Analyses are conducted on motion captured data during natural dialogue. Analysis results showed that nods frequently occur during speech utterances, not only for expressing dialog acts such as agreement and affirmation, but also as indicative of syntactic or semantic units, appearing at the last syllable of the phrases, in strong phrase boundaries. The paper also analyzes the dependence on linguistic, prosodic and voice quality information of other head motions, like shakes and tilts, and discuss about the potentiality for their use in automatic generation of head motions.

**Index Terms:** head motions, paralinguistic information, dialog acts, prosody, voice quality.

## 1. Introduction

Head motions often occur during speech utterances. Sometimes these motions are intentional and have meanings in communication, for example, nods are frequently used for expressing agreement, while shakes are used for expressing disagreement. However, most of time, the head motions are unconsciously produced. One of our motivations for the present work is to obtain a method to generate head motions from the speech signal, for example in an application of teleoperation of a humanoid robot (such as an android), where head motions would be automatically controlled from the operator's voice.

Many works tried to find a correspondence between head motions and prosodic features, such as the fundamental frequency (F0) contours which represents pitch movements.

For example, in [1], head motions were associated with speech over the fundamental frequency (F0). Experiments using read speech utterances of one American English speaker (ES) and one Japanese speaker (JS) showed following results for estimation of head motion from F0 and vice versa. From head motion to F0 the mean average was 73% for JS and 88% for ES. Opposite estimation from F0 to head motion showed a less obvious correlation. 25% mean average for JS and 50% for ES. In addition, correlation among F0 and the 6DOF, rotation and translation, of head motion for ES was between 39% and 52%, for JS between 22% and 30%, which is in average less than 50%. This shows that the only use of prosodic information is not enough to generate head motions.

In [2], experiments with one native speaker of Canadian English were carried out to enquire the correlation between speech (keywords like "left", "right" and "straight") and gestures (hand and head gestures). With focus on head nods and tilts a correspondence about ~64% between these two head motions and pitch accents have been observed. Through disregarding the phrase initial syllables, because they are a

known problem in prosody labeling, an increasing near 80% correspondence have been obtained. However, here it was not obvious when in the pitch a nod or a tilt occurs.

[3] reports that head motions are correlated with pitch and amplitude of the talker's voice, in Japanese read speech utterances, regarding all 6 DOF. For several sentences correlations were almost always over 50%, on average about 63%. Furthermore, in an experiment with animations, it is reported that a better perception of syllables have been achieved in a speech-in-noise task, with the normal, natural head motion compared with speech without head motion, double head motion and only auditory stimulus.

Variations in speech depending on head movement were observed in [4], for English sentences. Emphasis of a word often goes along with head nodding or a rise of the head can correspond with a rise in the voice. They call these movements 'visual prosody'. A talking head with these 'visual prosody' especially with head motion looks more natural even if this motion is not really connected with the content of the spoken text.

In [5], facial parameters (including head motions) were analyzed for short read Swedish utterances in which focal accent was systematically varied, in a variety of expressive modes including certain, confirming, questioning, uncertain, happy, angry and neutral. Results indicated that in all expressive modes, words with focal accent are accompanied by a greater variation of the facial parameters than the words in non-focal positions.

[6] compared head motion of neutral and emotional speech and found out that head motion is different depending on the emotional state. They investigated head motions for four emotional states, neutral, sadness, happiness and anger. As prosodic features, they used the pitch (F0), the RMS (root mean square) energy and their 1st and 2nd derivatives. For the head motion they took account of the 3 DOF of head rotation. Correlations between the prosodic features and head motion were about 74% for neutral emotional state and 69% for anger.

In [7], relations between head movements and the semantics of utterances are analyzed in Japanese spoken dialogue, with the purpose of improving spoken dialogue understanding by also using visual information. They consider the speaking turn, and speech functions.

We consider that the relationship between pitch and head motions can be language-dependent, since the function of the prosodic features may differ if for example the language is a tonal language (as in Chinese and Thai) or a lexical-accent language (as in Japanese).

Finally, although most of works dealing with head motion analysis study only features when a head motion occur, in the present work, we also analyze the features of speech utterances where head motions don't occur.

## 2. Data collection and annotation procedure

### 2.1. Data

About 30 minutes of free dialog conversations between two graduate students who know to each other were recorded. The target speaker is a female speaker, while the interlocutor is a male speaker. Simultaneous recordings of audio, video and motion capture data were conducted for the target speaker.

The motion capture system used is a Hawk system from Motion Analysis. Four cameras are arranged in an arc in front of the speaker. Thirty-seven (37) hemispherical passive reflective markers are applied to the speaker's face, head and upper-body. Although many markers were also placed on the face to capture lip motions and eye blinks, the 10 markers (placed on the head, nose, earlobes, chest and shoulders) shown in Fig. 1 were effectively used for characterizing the head motions. The markers on the head, nose and earlobes provide something of a static reference frame for the head, while the markers on the chest and shoulders provide a static reference for the upper body.

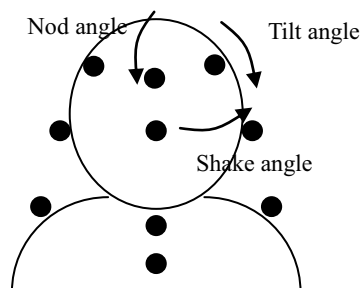


Figure 1: Markers and angles used to describe head motions.

The upper body motions given by the chest and shoulder markers are removed from the head markers based on singular value decomposition method [8], in order to obtain the head motions. The head rotation angles, used for describing the head motions, are also shown in Fig. 1.

### 2.2. Head motion tags

The following tag set was used to annotate the head motions, based on the display of the three angles along the time, and the video information.

- **fu (face up)**: the face is moved up.
- **fd (face down)**: the face is moved down.
- **nd (nod)**: single nod (down-up motion).
- **nm (multiple nods)**: multiple nods occur along the phrase.
- **ud (up-down)**: single up-down motion.
- **sh (shake)**: shakes (left-right motions) occur within the phrase.
- **ti (tilt)**: head tilts occur within the phrase.
- **no**: no head motions.

Nods were not necessarily realized by perfect vertical head motions. They can eventually be accompanied by a slight head tilt. We then consider the motion which has the strongest magnitude.

### 2.3. Linguistic information

A preliminary observation of the head motions in the data indicated that nods frequently occurred in particles (such as “*ne*”, “*de*”, “*kara*”). Interjections (such as “*un*”, “*ee*”, “*hee*”) were also often accompanied by a head motion. In order to verify the relations between such morphemes and the head motions, we decided to segment the utterances in phrase units (“*bunsetsu*” in Japanese), since such morphemes usually appear in the boundary of the phrases.

The speech utterances are manually segmented in phrase units, and the last morpheme of each phrase is transcribed by one native speaker of Japanese.

The annotation process resulted in a segmentation of 535 phrases. Special labels are also annotated for laugh, and breathing.

### 2.4. Dialog act tags

We also consider possible relationship between head motions and dialog acts, such as affirmative or negative reaction, expression of emotions like surprise or unexpectedness, and turn-taking functions.

Dialog act tags are annotated for each phrase in the dataset, according to the following set, based on the tags proposed in [9] for turn-taking and dialog acts.

- **k (keep)**: the speaker is keeping the turn; a short pause or a clear pitch reset is accompanied at the phrase boundaries.
- **k2 (keep)**: phrase boundaries in the middle of an utterance (when no pause exists between phrases). It corresponds to a “weak” phrase boundary.
- **k3 (keep)**: the speaker lengthens the phrase final, indicating that he/she is thinking, but is keeping the turn (can or not be followed by a pause).
- **f1 (filler)**: the speaker is thinking or preparing the next utterance, e.g., “*uum*”, “*eee*”, “*eeetoo*”, “*anoo*” (“*uhmmm*”).
- **f2 (conjunctions)**: can be considered as non-lengthened fillers, e.g., “*dakara*”, “*jaa*”, “*dee*” (“I mean”, “so”).
- **g (give)**: the speaker finished talking and is giving the turn to the interlocutor.
- **q (question)**: the speaker is asking for a question or a confirmation to the interlocutor.
- **bc (backchannels)**: the speaker is producing backchannels (agreeable responses) while the interlocutor is talking, e.g., “*un*” usually accompanied by a fall pitch movement, “*hai*” (“*uhm*”, “*yeah*”).
- **su (admiration/surprise/unexpectedness)**: the speaker is producing a reaction (admiration, surprise) to the interlocutor's utterances, e.g., “*hee*”, “*uso!*”, “*a!*” (“*wow*”, “*really?*”).
- **dn (denial, negation)**: E.g., “*ie*”, and “*uun*” accompanied by a fall-rise pitch movement (“*no*”).

Dialog act tags are annotated for each phrase by one subject, and later checked/corrected by another subject.

### 2.5. Prosodic and voice quality features

Although automatic procedures could be conducted for extraction of the phrase final tones, such as the methods proposed in [10] and [11], in the present work, we hand-

annotated the prosodic and voice quality features for analysis purposes, according to the following tags.

- **rs (rise)**: rising tone.
- **fa (fall)**: falling tone (includes reset-fall tones).
- **fr (fall-rise)**: fall pitch movement followed by a rise movement.
- **hi (high)**: high pitch.
- **mi (mid)**: middle-height pitch.
- **lo (low)**: low pitch.
- **cr (creaky)**: creaky voice or vocal fry (a voice quality characterized by very large intervals between glottal excitation pulses), when F0 is lowered and cannot be reliably measured.
- **wh (whisper)**: whisper (absence of F0).

The tone tags were annotated for each phrase, by one subject with experience in prosody and voice quality annotation.

### 3. Relation between head motions and speech

#### 3.1. Head motions and morphemes

Regarding the relations between morphemes and head motions, analysis results firstly indicated that nods frequently occur at the boundary of phrases, regardless the morpheme at the phrase boundaries. It frequently occurs at the particles, since particles usually appear at the boundary of phrases, but it was not restricted to the particles. Rather, nods seem to be more related with the dialog act functions carried by the morphemes at the phrase boundaries, as will be shown in Section 3.2.

Although only a small number of shakes were observed in the data, they occurred in utterances expressing negation or rejection. In the current dataset, shakes appeared in the morpheme “*uun*” (accompanied by a fall-rise intonation), and in utterances ending with “*...nai*” which express negation.

#### 3.2. Head motions and dialog acts

Table 1 shows the relationship between the head motions and the dialog act functions.

Table 1. *Distribution of the dialog acts (rows) and the head motions (columns).*

		<b>nd</b>	<b>fd</b>	<b>ud</b>	<b>fu</b>	<b>ti</b>	<b>sh</b>	<b>nm</b>	<b>no</b>
	total	<b>142</b>	24	28	20	33	4	3	189
<b>k</b>	61	<b>35</b>	<b>5</b>	1	0	3	0	0	17
<b>k2</b>	137	2	2	6	6	11	0	0	<b>106</b>
<b>k3</b>	28	4	1	2	1	4	0	0	<b>16</b>
<b>f1</b>	15	3	1	0	1	2	0	0	<b>8</b>
<b>f2</b>	22	2	0	0	3	5	0	0	<b>12</b>
<b>g</b>	79	<b>29</b>	<b>11</b>	<b>12</b>	<b>2</b>	5	2	1	14
<b>q</b>	25	<b>9</b>	3	3	2	0	0	0	7
<b>bc</b>	71	<b>58</b>	1	1	1	0	0	2	7
<b>su</b>	12	0	0	<b>2</b>	<b>4</b>	3	0	0	1
<b>dn</b>	4	0	0	1	0	0	<b>2</b>	0	1

From Table 1, it can be noted that nod (**nd**) was the head motion type which occurred with most frequency. Firstly, an expected result was that nods were present in almost all

backchannels (**bc**). Nods were also frequently observed at the strong phrase boundaries (**k**, **g**, **q**), regardless the presence or the type of a particle at the phrase final.

A surprising result was that even in the questions (**q**), where phrase finals are usually accompanied by a rising intonation, nods were more frequent than up-down or face-up motions. This is one of the factors that contribute to reduce the correlation between pitch and head motions.

A particular result was observed in pre-pause phrases when the subject keeps the turn (**k**): nods (**nd**) frequently occurred, while upward motions (**ud**, **fu**) never occurred.

Nods occur with less frequency at “weak” phrase boundaries in the middle of an utterance (**k2**), and at phrase boundaries where the speaker is thinking or indicates that he/she didn’t finish to utter (**k3**, **f**, **f2**). In these dialog act categories, predominance of no head motions (**no**) is observed in Table 1.

Phrases expressing surprise/admiration/unexpectedness (**su**) are usually accompanied by upward (**fu**, **ud**) or tilt motions (**ti**). As the number of **su** phrases is small in the present dataset, more detailed analysis in a larger database would be necessary to verify these trends in the expression of different emotions.

Nods sometimes occurred in the beginning of the phrase (10 phrases removed from Table 1). This is thought to be a kind of signal to the interlocutor that the speaker will take the turn and start to utter. Face up motions also sometimes occur at the beginning of the phrases with the same purpose.

Regarding the motion shapes, we observed that nods are often accompanied by a small upward motion before the usual down-up motion.

Sequence of multiple nods (**nm**) occurs along the whole utterance when the speaker is expressing agreement. When multiple nods occur in a sequence of backchannels such as in “*un un un un*”, usually the first nod was larger than the others.

Only 4 shakes (**sh**) were observed in the data. They occurred in utterances expressing negation or rejection, and were more dependent on the linguistic content, as discussed in Section 3.1.

Finally, tilts (**ti**) occurred in almost all dialog acts, excluding backchannels (**bc**), questions (**q**) and denial (**dn**). Although the direction of tilts (right or left) was also annotated, no significant differences were observed between their distributions.

#### 3.3. Head motions and prosodic and voice quality features

Relationship between pitch and head motions exists, but their correlation is not high, as also pointed out in [1]. Japanese is a pitch accent language, so that many pitch movements occur within the utterances due to the lexical accents. As described in the previous sections, our analyses indicated that the head motions don’t occur at every pitch accent nucleus, but rather, occur more frequently at the phrase boundaries.

However, even for the phrase boundary tones, a straight relation between pitch movements (tones) and head motions could not be observed. Table 2 shows the distributions between head motions and phrase final tones. Roughly, we could say that falling tones (**fa**) and creaky voice (**cr**) are usually accompanied by nods (**nd**), while high and middle-height pitch tones (**hi**, **mi**) are usually not accompanied by any head motion (**no**).

Although a clear correspondence could not be found between tones and head motions, Table 3 shows a better correspondence between tones and dialog acts. Rising tones

(rs) basically appear in questions (q), falling tones (fa) appear frequently in turn-keeping functions (k) and backchannels (bc), high and middle-height tones are frequent in weak phrase boundaries (k2), and low-pitch, creaky and whisper are frequent in turn-giving functions (g). The use of the morpheme information along with these tone information would produce a better correspondence with the dialog act functions.

Table 2. Distribution of the head motions (rows) and the phrase final tones (columns).

	rs	Fa	fr	hi	mi	lo	cr	wh
total	25	108	3	26	205	14	58	13
nd	8	83	1	0	26	4	20	4
fd	2	4	0	1	5	2	7	2
ud	5	1	0	1	11	3	6	1
fu	3	2	0	4	11	0	1	1
ti	1	4	0	2	20	1	4	1
sh	0	0	2	0	1	0	1	0
nm	0	0	0	0	3	0	0	0
no	6	14	0	18	126	3	18	4

Table 3. Distribution of the dialog acts (rows) and the phrase final tones (columns).

	rs	fa	fr	hi	mi	lo	cr	wh
k	1	45	0	1	12	0	2	0
k2	0	1	0	12	110	2	11	1
k3	0	10	0	4	12	0	2	0
f1	0	3	0	0	11	0	1	0
f2	0	2	0	2	14	0	4	0
g	2	1	0	1	20	9	36	10
q	19	1	1	1	1	0	1	0
bc	0	45	0	0	19	3	1	2
su	2	0	0	5	5	0	0	0
dn	1	0	2	0	1	0	0	0

Finally, regarding relations between head motions and other voice qualities (not included in the tables above), possible relations were observed mainly in nods and tilts. In confident utterances, accompanied by a normal or more pressed voice quality, nods tend to be more frequent, while in non-confident utterances, usually accompanied by a more lax or breathy voice quality, nods tend to be of smaller magnitude, and tilts or no head motions become more frequent. A deeper study would be necessary to verify such trends.

### 3.4. Some rules for head motion generation

Here we summarize some rules for generating head motions from speech, based on some of the analysis results in the previous sections.

- backchannels (“un”, “hai”, ...) + {falling or mid-height tones}: nods with high percentage of occurrence
- strong phrase boundaries + {low-pitch, falling tones, creaky, whisper}: nods with high percentage of occurrence.
- weak phrase boundaries + falling tone: nods with less percentage of occurrence
- denial/rejection words (“uun” + fall-rise tone, “iie”, ...): shakes with high percentage of occurrence.
- questions (usually accompanied by a rising tone): select nods or face up motions.

## 4. Conclusions

With the aim of generating head motions from speech signal, analyses were conducted for verifying the relations between head motions and dialog acts, prosodic features and linguistic information. Among the several head motions, nods were the most frequent, appearing not only for expressing dialog acts such as agreement or affirmation, but also as indicative of syntactic or semantic units, appearing at the last syllable of the phrases, in strong phrase boundaries. Tilts usually occur at the utterances where the speaker is thinking or is not confident, and is possibly accompanied by a lax or breathy voice quality. Shakes are used to express negation/denial and are more dependent on the content of the speech utterance.

The next step of our work is to use the current analysis results, generate head motions from the information carried by speech, and evaluate its naturalness in a humanoid robot head. Future works include head motion analysis of other speakers and evaluation of individuality.

## 5. Acknowledgements

This research is supported by the Ministry of Internal Affairs and Communications. We thank Judith Hass for helping in the motion data collection. We also thank Kyoko Nakanishi for helping in the data analysis.

## 6. References

- [1] Yehia, H.C., Kuratate, T., Vatikiotis-Bateson, E. “Linking facial animation, head motion and speech acoustics,” *J. of Phonetics* 30, 555-568, 2002.
- [2] Sargin, M.E., Aran, O., Karpov, A., Ofli, F., Yasinnik, Y., Wilson, S., Erzin, E., Yemez, Y., Tekalp, A.M. “Combined Gesture-Speech Analysis and Speech Driven Gesture Synthesis,” *Proc IEEE International Conference on Multimedia*, 2006.
- [3] Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., Vatikiotis-Bateson, E. “Visual prosody and speech intelligibility – Head movement improves auditory speech perception,” *Psychological Science* 15 (2), 133-137, 2004.
- [4] Graf, H.P., Cosatto, E., Strom, V., Huang, F.J. “Visual prosody: Facial movements accompanying speech,” *Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition (FGR’02)*, 2002.
- [5] Beskow, J., Granstrom, B., House, D. “Visual correlates to prominence in several expressive modes,” *Proc. Interspeech 2006 – ICSLP*, 1272-1275, 2006.
- [6] Busso, C., Deng, Z., Grimm, M., Neumann U., Narayanan, S. “Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis,” *IEEE Trans. on Audio, Speech and Language Processing*, March 2007
- [7] Iwano, Y., Kageyama, S., Morikawa, E., Nakazato, S., Shirai, K. “Analysis of head movements and its role in spoken dialogue,” *Proc. ICSLP’96*, 2167-2170, 1996.
- [8] Stegmann, M.B., Gomez, D.D. “A brief introduction to statistical shape analysis,” published online, 2002.
- [9] Ishi, C.T., Ishiguro, H., Hagita, N. “Analysis of prosodic and linguistic cues of phrase finals for turn-taking and dialog acts,” *Proc. Interspeech’2006 - ICSLP*, 2006-2009, 2006.
- [10] Ishi, C.T. “Perceptually-related F0 parameters for automatic classification of phrase final tones,” *IEICE Trans. Inf. & Syst.*, E88-D(3), 481-488, 2005.
- [11] Ishi, C.T., Ishiguro, H., Hagita, N. “Using Prosodic and Voice Quality Features for Paralinguistic Information Extraction,” *CD-ROM Proc. Speech Prosody 2006*, 2006.