Evaluation of prosodic and voice quality features on automatic extraction of paralinguistic information

Carlos Toshinori Ishi[*], Hiroshi Ishiguro[*†], Norihiro Hagita[*]

[*] Intelligent Robotics and Communication Labs.
ATR
Kyoto, Japan

[†] Faculty of Engineering
Osaka University
Osaka, Japan

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

***Abstract* - Aiming to realize a non-verbal communication between humans and robots, the use of acoustic parameters related with voice quality features, besides classical prosodic features, is proposed and evaluated for automatic extraction of paralinguistic information (intentions, attitudes, and emotions) in dialog speech. Experimental results indicated that prosodic features were effective for detecting groups of paralinguistic information expressing specific functions (such as affirmation, denial, and asking for repetition), accounting for 61 % of the global identification rate. Voice quality features were effective for detecting part of the paralinguistic information expressing emotions or attitudes (such as surprise, disgust and admiration), leading to 12 % improvement in the global identification rate.**

Index Terms – Prosody, voice quality, paralinguistic information, non-verbal communication, automatic detection.

I. INTRODUCTION

Recent works in communication robots have paid attention to non-verbal communication processing. In robot environments, there are cases where linguistic information recognition fails, but non-verbal information can be extracted, so that a communication robot could be constructed, even with the use of only non-verbal information.

The information carried by speech in communication, can be categorized as linguistic (verbal) and paralinguistic (non-verbal). Linguistic information recognition would be a powerful function in interactive robots. However, the performance of current speech recognition technologies, which are focused on linguistic information extraction, is restricted by many factors, such as intra-speaker, inter-speaker, environment and context variabilities. As linguistic information is a critical part of conversation, any failures in its recognition would give a negative impression, leading humans to be disappointed in interacting with robots. Thus, we consider that the current technology for linguistic information recognition is not currently appropriate for interactive robots.

To complement linguistic information recognition, non-verbal information extraction could be used as an alternative technology. Non-verbal information includes gestures and paralinguistic information. Fig. 1 shows the block diagram of a communication robot regarding both verbal and non-verbal information. Currently, there are many works related with gestures in human-robot interaction (e.g. [1]). However, there are only a few, related with the use of paralinguistic information in robot communication [2]. In the present research, we focus on paralinguistic information extraction, having as a goal, a robot that can keep interaction with humans, even in the cases where speech recognition fails.

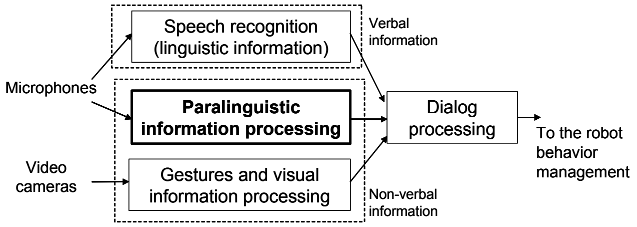
Speech recognition (linguistic information)
Verbal information
Microphones
Paralinguistic information processing
Dialog processing
To the robot behavior management
Video cameras
Gestures and visual information processing
Non-verbal information

Fig. 1 Block diagram of a communication robot system considering verbal and non-verbal information.

The understanding of paralinguistic information becomes as important as linguistic information, especially in non-verbal communication using grunt-like utterances, such as “eh”, “ah”, and “un”. Such utterances are frequently used to express a reaction to the interlocutor’s utterance in a dialog scenario, and usually express some intention, attitude, or emotion. As there is little phonetic information represented by such grunt-like utterances, most of the paralinguistic information is likely represented by variations in prosodic or voice quality features.

Up till now, most works dealing with paralinguistic information extraction have focused only on prosodic features like fundamental frequency (F0), power and duration. However, when analysing natural conversational speech data, the presence of several voice qualities (caused by non-modal phonation, such as breathy, whispery, creaky and harsh [3]) is often observed, mainly in expressive speech utterances [4]. For example, whispery and breathy voices are reported to correlate with the perception of fear [5], sadness, relaxation and intimate in English [6], and politeness in Japanese [7]. Vocal fry or creaky voices appear in low tension voices correlating with sad, bored or relaxed voices [5,6], or in pressed voices expressing admiration or suffer [8]. Harsh and ventricular voices are reported to correlate with anger, happiness and stress [5,6].

Further, in segments uttered by such voice qualities (caused by non-modal phonation types), F0 information is often missed by F0 extraction algorithms due to the irregular characteristics of the vocal fold vibrations. Therefore, in such segments, the only use of prosodic features would not be enough for its complete characterization. Thus, other acoustic features related with voice quality become important for a more suitable characterization of the speaking style.

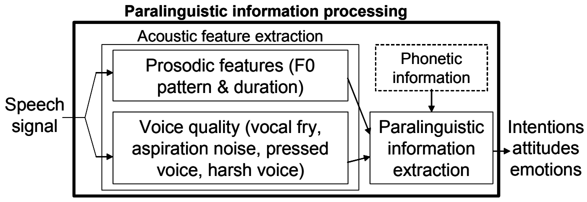
Paralinguistic information processing
Acoustic feature extraction
Prosodic features (F0 pattern & duration)
Voice quality (vocal fry, aspiration noise, pressed voice, harsh voice)
Phonetic information
Paralinguistic information extraction
Speech signal
Intentions attitudes emotions

Fig. 2 Block diagram of the proposed framework for paralinguistic information extraction.

Fig. 2 shows the framework proposed for extraction of paralinguistic information, by using information of voice quality features, besides the classical prosodic features. In previous works, we have proposed several acoustic parameters for representing the features of intonation and specific voice qualities [9-11]. In the present work, we extend and improve these acoustic parameters, and evaluate their performance in the automatic extraction of paralinguistic information.

II. Description of the speech data for analysis and experimental setup

The utterances “e” and “un” (including variations such as “e”, “eh”, “ee”, “eeee”, “hee”, and “un”, “nnn”, “uhn”, etc.) are chosen here for analysis, because they are often used to express a reaction in Japanese conversational speech, and carry a large variety of paralinguistic information (**PI**) depending on its speaking style. Possible PI (speech acts, attitudes or emotions) transmitted by varying the speaking styles of the utterances “e” and “un” are listed in Table I.

TABLE I
List of paralinguistic information carried by “e” and “un”

Paralinguistic information (PI)	Abreviation
affirmation	***aff***
agreement, understanding, consent	***agr***
backchannel (agreeable responses)	***backch***
denial	***den***
filler (think)	***filler***
embarrassment, hesitation	***emb***
admiration	***adm***
envy	***env***
asking for a repetition	***askrep***
surprise, amazement, astonishment	***surp***
unexpectedness	***unexp***
suspicion	***susp***
blame, criticism	***blm***
disgust, dislike	***disg***
dissatisfaction	***dissat***

The list of table I was obtained by referring to the list of speech acts annotated for the utterances “e” and “un” in the JST/CREST ESP Expressive Speech Database [12]. The items of the list have been obtained by free-text annotations of 4 subjects, in “e” and “un” utterances appearing in natural conversations of the database. The annotated words have been arranged by the 4 subjects for reducing redundancies. We do not guarantee that this list contains all the possible PI the utterances “e” and “un” can carry. However, we considered that this list is rich enough for our purposes of human-robot communication.

Here, speech data is newly recorded in order to get a balance in terms of the PI carried by the utterance “e” or “un”. For that purpose, sentences are elaborated in such a way to induce the subject to produce a specific PI. Two sentences are elaborated for each PI item of Table I.

The sentences are first read by one native speaker. These sentences will be referred as “inducing utterances”. Then, subjects are asked to produce a target reaction, i.e., utter in a way to express a determined PI, through the utterance “e”, after listening to each pre-recorded inducing utterance. The same procedure was conducted for the utterance “un”. Some short sentences are also elaborated to be spoken after the utterance “e” or “un”, in order to get a reaction as natural as possible. However, a pause is requested between the utterance “e”/”un” and the following short utterance. Further, the utterance “he” (with the aspirated consonant /h/ before the vowel /e/) is allowed to be spoken, if the subject judges that it is more appropriated for expressing some PI.

Utterances spoken by 6 subjects (2 male and 4 female speakers between 15 to 35 years old) are used for analysis and evaluation. In addition to the PI list, speakers are also asked to utter “e”, “he” and “un” in a pressed voice quality, which frequently occurs in natural expressive speech [8], but was found more difficult to naturally occur in an acted scenario.

All the utterances “e” and “un” are manually segmented for subsequent analysis and evaluation.

III. Perceptual voice quality labels and relationship with Paralinguistic Information

Perceptual voice quality labels are annotated for two reasons. One is to verify their effects on the expression of different PI. Another is to use them as targets for evaluating the automatic detection of voice qualities. The perceptual voice quality labels are annotated by one subject with experience in voice quality (the author oneself), according to the following criteria.

- *w*: strong aspiration noise is perceived along the utterance.
- *a*: strong aspiration noise is perceived in the syllable offset.
- *h*: harsh voice (rasping sound, aperiodic noise) is perceived.
- *c*: vocal fry or creaky voice is perceived.
- *p*: pressed voice is perceived.

A question mark “*?*” was added for each voice quality label, if their perception is not clear. Fig. 3 and 4 show the distributions of the perceived voice quality categories for each PI item.

In Fig. 3, we can first observe that soft aspiration noise (*w?*) is perceived in some utterances of almost all PI items. In contrast, strong aspiration noise (*w*), harsh or harsh whispery voices (*h, hw*) and syllable offset aspiration noise (*a, a?*) are perceived in PI items expressing some emotion or attitude (admiration, surprise, unexpectedness, suspicion, blame, disgust and dissatisfaction). This indicates that the detection of these voice qualities (*w, h, hw, a*) could be useful for the identification of these expressive PI items.

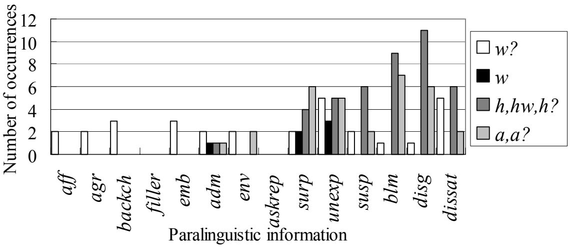

Fig. 3 Distribution of perceived categories of aspiration noise and harsh voices, for each paralinguistic information item.

In Fig. 4, we can observe that creaky voices are perceived in *filler*, *emb*, *adm* and *disg*. However, the additional perception of pressed voices is important to discriminate between emotionless fillers, and utterances expressing some emotion or attitude (admiration, disgust and embarrassment).

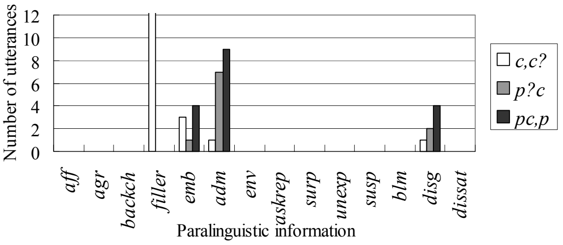

Fig. 4 Distribution of perceived categories of creaky voice and pressed voice, for each paralinguistic information item.

IV. Acoustic parameters

In this section, we describe acoustic parameters that potentially represent the perception of prosodic and voice quality features, which are responsible for the discrimination of different speaking styles, and verify their performance in automatic detection.

A. Acoustic parameters related with prosodic features: F0move and duration

In [9], a set of parameters was proposed for describing the intonation of phrase finals (phrase final syllables), based on F0 and duration information. Here, we use a similar set of parameters with some modifications, for the monosyllabic "e" and "un" utterances.

For the pitch-related parameters, F0 is first estimated based on the normalized autocorrelation function of the LPC inverse-filtered residue of the pre-emphasized speech signal. Details about the F0 estimation procedure can be found in [9]. All F0 values are converted to the musical (log) scale before any subsequent processing. The expression (1) shows a formula to produce F0 in semitone intervals.

$$F0[semitone] = 12 * \log_2(F0[Hz]) \qquad (1)$$

In [9], each syllable is broken in two segments of equal length, and representative F0 values are extracted for each segment. Several candidates for the representative F0 values have been tested in [9]. Here, we use the ones that best matched with perceptual scores of the F0 movements. For the first segment, an average value is estimated using F0 values within the segment (*F0avg2a*). And for the second segment, a target value is estimated as the F0 value at the end of the segment of a first order regression line of F0 values within the segment (*F0tgt2b*). A variable called ***F0move*** is then defined as the difference between *F0tgt2b* and *F0avg2a*, quantifying the amount and direction of F0 movement within the syllable. *F0move* is positive for rising F0 movements, and negative for falling movements. Details about the evaluation of these parameters can be found in [9].

The representation of F0 movements by *F0move* is valid when F0 only rises, only falls, or does not change within a syllable. This condition is true for most cases in Japanese syllables. However, there are cases where F0 falls down and then rises up within the same syllable. A fall-rise intonation is commonly used in "un" utterances for expressing a denial.

In the present work, we proposed a method for detecting fall-rise movements, by searching for negative F0 slopes in the syllable nucleus, and positive F0 slopes in the syllable end portion. Here, syllable nucleus is defined as the 25 % to 75 % center portion of the syllable duration, while the syllable end is defined as the 40 % to 90 % portion of the syllable. The initial and final portions of the syllable are removed from the slope searching procedure, in order to avoid misdetection of F0 movements due to co-articulation effects.

If a fall-rise movement is detected, the syllable is divided in three portions of equal length. The representative F0 value of the first portion is estimated as the average F0 value (*F0avg3a*). For the second portion, the minimum F0 value (*F0min3b*) is estimated. Finally, for the last portion, a target value (*F0tgt3c*) is estimated in the same way of *F0tgt2b*. Then, two *F0move* values are estimated. *F0move = F0min3b – F0avg3a*, representing the falling degree, and *F0move = F0tgt3c – F0min3b*, representing the rising degree.

Fall-rise tones were correctly detected in all "un" utterances expressing denial. It was also detected in two "e" utterances. However, in these two cases, the *F0move* values of the falling movement were smaller than 2 semitones, indicating a slight falling movement. In contrast, the *F0move* values for the "un" utterances expressing denial were all larger than 3 semitones.

For utterance ***duration***, the manually segmented boundaries could be straightly used, since the utterances are monosyllabic. However, as the manual segmentation may contain some silence (non-speech) portions close to the segmentation boundaries, an automatic procedure is further conducted, by estimating the maximum power of the syllable, and moving the boundaries until the power becomes 20 dB weaker than the maximum power. The newly segmented boundary intervals are used as segmental duration.

B. Detection of vocal fry (creaky voice): PPw, IFP, IPS

Creaky voice or vocal fry is characterized by the perception of very low fundamental frequencies, where individual glottal pulses can be heard, or by a rough quality caused by an alternation in amplitude, duration or shape of successive glottal pulses.

Here, we use the algorithm proposed in [10] for detection of vocal fry segments. A simplified block diagram of the detection algorithm is shown in Fig. 5. The algorithm first searches for power peaks in a "very short-term" power contour (obtained by using 4 ms frame length each 2 ms), which reflects the impulse-like properties in very low fundamental frequencies, characteristic of vocal fry signals. Then, it checks for constraints of periodicity and similarity between successive glottal pulses. The algorithm depends basically on three parameters: power thresholds for detection of power peaks (*PPw*), intra-frame periodicity (*IFP*), which is based on the normalized autocorrelation function, and inter-pulse similarity (*IPS*), which is estimated as a cross-correlation between the speech signals around the detected peaks. Here, vocal fry segments are detected by using *PPw* larger than 7 dB, *IFP* smaller than 0.8, and *IPS* larger than 0.6. Details about the parameters can be found in [10].

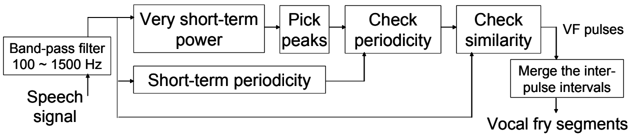

Fig. 5 Simplified block diagram of the vocal fry detection.

C. Detection of pressed voice: H1'-A1'

Lax and pressed voices are both present in the creaky voice utterances detected by the method described in the previous section. Lax creaky voices appear in relaxed voices indicating boredom or sadness [6,7]. On the other hand, pressed creaky voices indicate strong attitudes/feelings of admiration or disgust [8]. Therefore, detection of pressed voices is important for PI discrimination.

The production mechanism of pressed voice is not clearly explained yet, but it possibly has features similar to "tense voice" [13]. A difference between "tense voice" and "lax voice" is reported to appear in the spectral tilt, since in tense voice, the glottal excitations become more impulse-like, and the higher frequency components are emphasized in relation to the fundamental frequency component. Acoustic parameters like *H1-H2* and *H1-A1* are proposed to reflect the effects of spectral tilt [13]. *H1* is the amplitude power of the first harmonic (fundamental frequency), *H2* is the amplitude power of the second harmonic, and *A1* is the amplitude power of the harmonic closest to the first formant.

Here, we take these acoustic parameters into account. However, in creaky or harsh voices, the irregularities in periodicity cause disturbances in the harmonic structure of their spectrum, so that it becomes difficult to extract harmonic components from the spectrum. In the present work, if periodicity is not detected, instead of *H1*, we use the maximum peak power of a low frequency band of 100 to 200 Hz (*H1'*). Also, as an automatic formant extraction is difficult, instead of *A1*, we use the maximum peak power in the frequency band of 200 to 1200 Hz (*A1'*), where the first formant is likely to appear. If periodicity is detected, *H1'* is equalized to *H1*. Preliminary experiments indicates pressed voice can be detected, when *H1'-A1'* is smaller than -15 dB, for each frame.

D. Detection of aspiration noise: F1F3syn, A1-A3

Aspiration noise refers to turbulent noise due to an air escape at the glottis, occurring in whispery and breathy voices. Although there is a distinction between whispery and breathy voices from a physiological viewpoint [3], a categorical classification of voices in whispery or breathy is difficult in both acoustic and perceptual spaces [14]. Further, aspiration noise is also often perceived in harsh voices, which is called harsh whispery voice in [3]. In the present work, we use a degree of aspiration noise as indicative of such voice qualities.

The aspiration noise detection algorithm is based on the proposed in [11]. The algorithm depends basically on two parameters, shown in Fig. 6. The main parameter, called *F1F3syn*, is a measure of synchronization (using a cross-correlation measure) between the amplitude envelopes of the signals obtained by filtering the input speech signal in two frequency bands, one around the first formant (F1) and another around the third formant (F3). If aspiration noise is absent, *F1F3syn* has values close to 1, while if it is present, *F1F3syn* has values closer to 0. The second parameter, called *A1-A3*, is a measure of the difference (in dB) between the powers of F1 and F3 bands. This parameter is used to constraint the validity of the *F1F3syn* measure, when the power of F3 band is too lower than that of F1 band, so that aspiration noise could not be clearly perceived. F1 band is set to 100 ~ 1500 Hz, while F3 band is set to 1800 ~ 4500 Hz. More details about the evaluation of the method can be found in [11]. Here, aspiration noise is detected when *F1F3syn* is smaller than 0.4 and *A1-A3* is smaller than 25 dB.

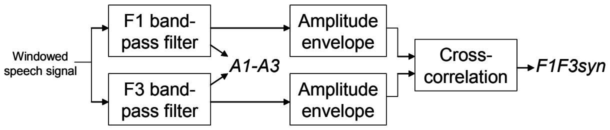

Fig. 6 Simplified block diagram of the acoustic parameters for aspiration noise detection.

E. Detection of harsh and/or whispery voices

As shown in Fig. 3, there is no clear distinction in functionality between harsh, harsh whispery and whispery voices (*h, hw, w, a*). Thus, all these voice qualities will be treated as one category, hereinafter.

The aperiodicity, characteristic of harsh voices, is here detected when neither periodicity nor creaky voice is detected. Also, the initial and final 3 frames of each utterance are eliminated, for avoiding the effects of F0 disturbances at the onset and offset of the syllables.

F. Evaluation of automatic detection of voice qualities

Fig. 7 shows a summary of the results for automatic detection of the voice qualities discussed in the previous sections.

The detection of creaky voice is evaluated by an index called *VFR* (Vocal Fry Rate), defined as the duration of the segment detected as vocal fry (*VFdur*) divided by the total duration of the utterance. Fig. 7 shows the results of detection of creaky segments, by using a criterion of *VFR* > 0.1. We can note that all creaky segments are correctly detected (about 90% for *c, c?*), with only a few insertions (*non c*).

For evaluating pressed voice detection, an index called *PVR* (Pressed Voice Rate) is defined as the ratio between the duration of the segment detected as pressed (*PVdur*), by the utterance duration. An utterance is detected as pressed, if *PVR* is larger than 0.1, and *PVdur* is larger than 100 ms, indicating that the segment has to be long enough to be perceived as pressed. 69 % of the pressed voice utterances were correctly identified in (*p,pc, p?*). Among them, most "e" utterances were correctly identified, while the detection failed in most of "un" utterances. This is probably because the nasal formant (around 100 to 300 Hz) increases the spectral power in the lower frequencies, consequently rising the *H1'-A1'* value. More robust acoustic features have to be investigated for detecting pressed voice in nasalized vowels.

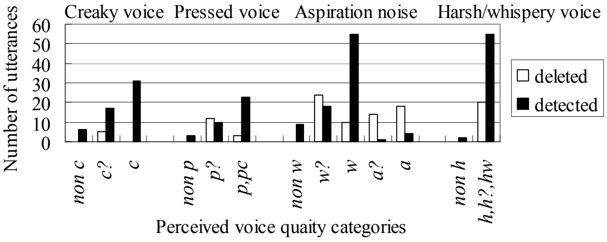

Fig. 7 Results of automatic detection of voice qualities, for each perceived category.

As in the previous voice qualities, an index called *ANR* (Aspiration Noise Rate) is defined as the duration of the segment detected as aspirated (*ANdur*) divided by the total duration of the utterance. Utterances containing aspiration noise are detected by using a criterion of *ANR* > 0.1. Most of the utterances where strong aspiration noise was perceived (*w*) could be correctly detected (81%). However only a few utterances could be detected by using *ANR*, where aspiration noise was perceived in the syllable offset (*a?* and *a*), as shown in Fig. 7. This is because these syllable offset aspirations are usually unvoiced, and very short in duration. Other methods have to be evaluated for the detection of such aspirated syllables.

Finally, an index called *HWR* (Harsh Whispery Rate) is defined as the summation of *APdur* (duration of the segment detected as aperiodic) and *ANdur*, divided by the utterance duration. 73 % of the utterances perceived as harsh and/or whispery (*h,h?,hw*) could be detected by using *HWR* > 0.1, and only a few insertion errors were obtained (*non h*), as shown in Fig. 7.

V. Identification of Paralinguistic Information Based on Prosodic and Voice Quality Features

In 32 of the total of 363 utterances, *F0move* could not be estimated due to missing F0 values. These missing values are due to non-modal phonations causing irregularities in the periodicity of the vocal folds. Fig. 8 shows the distributions of the prosodic features (*F0move* vs. *duration*), excluding the utterances where *F0move* could not be obtained due to missing F0 values in non-modal phonations, and the ones where fall-rise intonation was detected.

Thresholds for *F0move* and *duration* are set, based on a preliminary evaluation of classification trees for discriminating the present PI. A threshold of -3 semitones is set for *F0move* to discriminate falling tones (*Fa*), while a threshold of 1 semitone is set for rising tones (*Rs*). Utterances where *F0move* is between -3 and 1 semitone are considered as flat tones (*Ft*). The 32 utterances, where F0move could not be obtained, are also treated as flat tones in the evaluation of automatic detection. Two thresholds are also set for *duration*. Utterances shorter than 0.36 seconds will be called short (*S*), while utterances with duration between 0.36 and 0.6 seconds will be called long (*L*). Utterances longer than 0.6 seconds will be called extremely long (*E*).

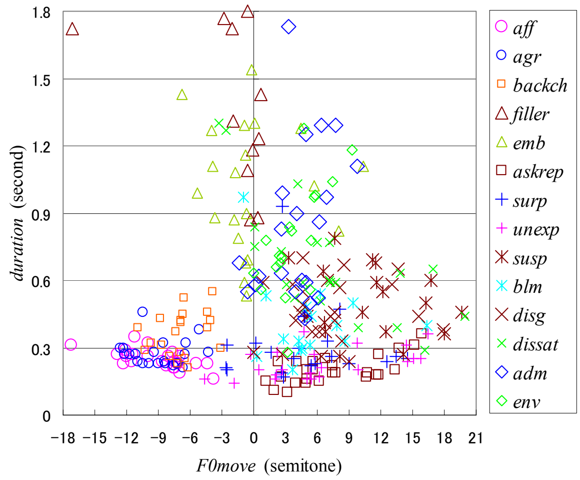

Fig. 8 Distribution of the prosodic features (*F0move* vs. *duration*) for each PI.

Table II show the results for automatic detection of PI items based on the prosodic and voice quality parameters described in Section IV. The automatic identification of all PI items is difficult since many PI items share the same speaking styles. For example, there is little or no distinction in speaking style between affirmation and agreement, or between surprise and unexpectedness. Thus, the PI items which share similar speaking styles and which carry similar meanings in communication, are grouped for evaluating the automatic detection. For example, affirmation, agreement and backchannel are positive reactions, while suspicion, blame, disgust and dissatisfaction, are all negative reactions. The resulting seven PI groups are separated by horizontal lines in Table II. The detection rate in the last column of Table II is calculated for each PI item, by counting the utterances belonging to their most representative speaking style.

Results indicate detection rates higher than 90 %, for positive reactions (*aff, agr, backch*), denial, filler and asking for repetition. Among the positive reactions, affirmation tends to be uttered by short fall intonation (*SFa*), while longer utterances (*LFa*) are more likely to appear when expressing agreement and backchannels. The fall-rise tone (*FaRs*)

TABLE II

AUTOMATIC DETECTION OF PARALINGUISTIC INFORMATION BASED ON PROSODIC AND VOICE QUALITY FEATURES. NUMBERS WITHIN PARENTHESIS INDICATE UTTERANCES WHERE HARSH AND/OR WHISPERY VOICE IS DETECTED (*HWR* > 0.1). PV IDICATES DETECTED PRESSED VOICE UTTERANCES (*PVR* > 0.1).

	total	*FaRs*	*SFa*	*LFa*	*EFa*	*SFt*	*LFt*	*EFt*	*SRs*	*LRs*	*ERs*	*PV*	Detection rate
aff	25		***24***			1							96 %
agr	22		***20***	***2***									100 %
backch	20		***9***	***11***									100 %
den	15	***15***											100 %
filler	18				***5***			***13***					100 %
emb	31				***7***		***2***	***12***			7	3	67 %
askrep	26								***24*** (1)	1			92 %
surp	23					***4 (1)***	0 (1)		9 ***(4)***	0 (3)	1		39 %
unexp	23		1			***3 (1)***			11 ***(5)***	1 (1)			39 %
susp	28					0 (1)	0 ***(1)***	0 ***(1)***	3 (2)	***11 (2)***	7		54 %
blm	23					1 (3)	0 ***(2)***	1 ***(2)***	3 (4)	***6 (1)***			48 %
disg	31					1	1	2***(1)***	4	***9 (5)***	5	3	58 %
dissat	24				1	1		3 ***(1)***	1	***8 (2)***	5 (2)		46 %
adm	33				1		2			4 (1)	***11 (1)***	***13***	76 %
env	21						1	1	1	4 (1)	***13***		62 %

detection was enough for the identification of denial. Extremely long fall or flat tones (*EFa, EFt*) were effective to identify fillers.

Short rise tones (*SRs*) identifies asking for repetition, surprise, or unexpectedness. Part of the utterances in *SRs* could be identified as *surp/unexp* by the detection of strong aspiration noise or harshness (numbers within the parenthesis in Table II). However, part of the utterances in *surp/unexp* has similar speaking styles with *askrep*. In these cases, the context has to be taken into account for discrimination.

The big overlap in the rising tones shown in Fig. 8, resulted in lower detection rates for *surp/unexp*, *susp/blm/ disg/dissat*, and *adm/env*, as shown in the bottom half of Table II. Although harsh and/or whispery voice quality is more indicative of negative reactions, rather than admiration or envy, other acoustic parameters related with brightness could be useful for discriminating them. Most of the detection errors in the rising tones are thought to be due to context dependency. However, part of these detection errors could be reduced, by improving the detection of voice quality features.

Pressed voice detection (*PV*) was effective for identifying part of utterances expressing admiration. The discrimination of *PV* utterances appearing in *disg* and *emb* could need context information. However, it was observed that most utterances in *adm* are "he", while most utterances in *disg* and *emb* are "e".

The overall detection rate using these simple thresholds for discrimination of the seven PI groups shown in Table II was 73 %. Regarding the contribution of the acoustic parameters used in the present work, 61 % of the correct identification was due to the only use of prosodic features, while 12 % was due to voice quality parameters.

VI. CONCLUSION

We proposed and evaluated the use of prosodic and voice quality features for automatic extraction of paralinguistic information in dialog speech. We showed that prosodic features are effective to detect paralinguistic information items expressing some functions, such as affirmation, denial, filler, and asking for repetition. Voice quality features were shown to be effective for identifying part of paralinguistic information items expressing some emotion or attitude (surprise/ unexpectedness, suspicion/blame/disgust/ dissatisfaction, and admiration/envy).

Improvements in the detection of voice qualities (harshness, pressed voices in nasalized voices, and syllable offset aspiration noise) can still improve the detection rate of paralinguistic information items expressing emotions/attitudes.

Future works are improvement of voice quality detection, investigations about how to deal with context information, and evaluation in a human-robot interaction scenario.

ACKNOWLEDGEMENT

This research was supported by the Ministry of Internal Affairs and Communications.

REFERENCES

[1] T. Kanda, H. Ishiguro, M. Imai, and T. Ono, "Body Movement Analysis of Human-Robot Interaction," *International Joint Conference on Artificial Intelligence* (*IJCAI 2003*), pp.177-182, 2003.

[2] S. Fujie, Y. Ejiri, Y. Matsusaka, H. Kikuchi, and T. Kobayashi, "Recognition of paralinguistic information and its application to spoken dialogue system," *IEEE Workshop on Automatic Speech Recognition and Understanding* (*ASRU '03*), pp. 231-236, 2003.

[3] J. Laver, "Phonatory settings," In *The phonetic description of voice quality*. Cambridge University Press, 1980, pp. 93-135.

[4] D. Erickson, "Expressive speech: production, perception and application to speech synthesis," *Acoust. Sci. & Tech.*, vol. 26, no. 4, pp. 317-325, 2005.

[5] G. Klasmeyer and W.F. Sendlmeier, "Voice and Emotional States," In *Voice Quality Measurement*, Singular Thomson Learning, 2000, pp. 339-358.

[6] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication* 40, pp. 189-212, 2003.

[7] M. Ito, "Politeness and voice quality – The alternative method to measure aspiration noise," Proc. *Speech Prosody 2004*, pp. 213-216, March 2004.

[8] T. Sadanobu, "A natural history of Japanese pressed voice", *J. of Phonetic Society of Japan*, vol. 8, no. 1, pp. 29-44, 2004.

[9] C.T. Ishi, "Perceptually-related F0 parameters for automatic classification of phrase final tones," *IEICE Trans. Inf. & Syst.*, vol. E88-D, no. 3, pp. 481-488, 2005.

[10] C.T. Ishi, H. Ishiguro, and N. Hagita, "Proposal of acoustic measures for automatic detection of vocal fry," Proc. *Eurospeech 2005*, pp. 481-484, 2005.

[11] C.T. Ishi, "A new acoustic measure for aspiration noise detection," Proc. *ICSLP 2004*, vol. II, pp. 941-944, 2004.

[12] http://feast.atr.jp/esp/esp-web/

[13] M. Gordon, P. Ladefoged, "Phonation types: a cross-linguistic overview," *J. of Phonetics* 29, pp. 383-406, 2001.

[14] J. Kreiman and B. Gerratt, "Measuring vocal quality," In *Voice Quality Measurement*, Singular Thomson Learning, 2000, pp. 73-102.