Using Prosodic and Voice Quality Features for Paralinguistic Information Extraction

Carlos Toshinori Ishi, Hiroshi Ishiguro & Norihiro Hagita

Intelligent Robotics and Communication Laboratories
ATR, Kyoto, Japan
{carlos; hagita}@atr.jp ishiguro@ed.ams.eng.osaka-u.ac.jp

Abstract

The use of voice quality features in addition to prosodic features is proposed for automatic extraction of paralinguistic information (like speech acts, attitudes and emotions) in dialog speech. Perceptual experiments and acoustic analysis are conducted for monosyllabic utterances spoken in several speaking styles, carrying a variety of paralinguistic information. Acoustic parameters related with prosodic and voice quality features potentially representing the variations in speaking styles are evaluated. Experimental results indicate that prosodic features are effective for identifying some groups of speech acts with specific functions, while voice quality features are useful for identifying utterances with an emotional or attitudinal expressivity.

1. Introduction

Besides the linguistic information, the understanding of paralinguistic information is also important in spoken dialog systems, 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 sort of intention, attitude, or emotion. Also, 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 analyzing natural conversational speech data, the presence of several voice qualities (caused by non-modal phonations, such as breathy, whispery, creaky and harsh [1]) is often observed, mainly in expressive speech utterances [2]. Whispery and breathy voices are characterized by the perception of a turbulent noise (aspiration noise) due to air escape at the glottis, and are correlated with the perception of fear [3], sadness, relaxation and intimate in English [4], and politeness in Japanese [5]. Vocal fry or creaky voices are 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. Vocal fry may appear in low tension voices correlating with sad, bored or relaxed voices [3,4], or in pressed voices expressing admiration or suffer [6]. Harsh and ventricular voices are characterized by the perception of an unpleasant, rasping sound, caused by irregularities of vocal fold vibrations in higher fundamental frequencies, and are reported to correlate with anger, happiness and stress [3,4].

Further, in segments uttered by such voice qualities (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 a complete characterization of the segment. Thus, other acoustic features related with voice quality become important for a more suitable characterization of the speaking style.

In previous researches, we have proposed several acoustic parameters for representing the features of specific voice qualities [7-10]. Here, we propose a framework for extraction of paralinguistic information as shown in Figure 1, using voice quality related acoustic parameters, in addition to prosodic features.

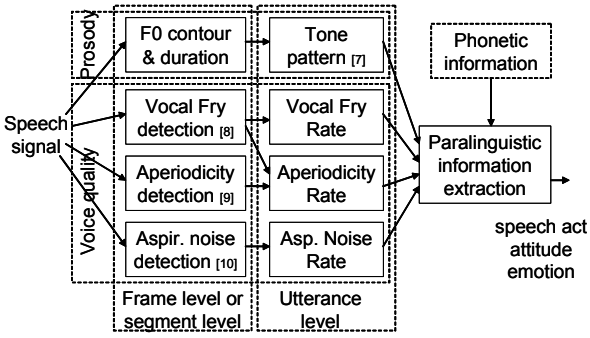

Figure 1: *Block diagram of the proposed framework for paralinguistic information extraction.*

2. Description of the speech data for analysis

The utterance “e” (including variations such as “e”, “eh”, “ee”, “eeee”, “hee”, etc.) is chosen here for analysis, because it is often used to express a reaction in Japanese conversational speech, and carries a large variety of paralinguistic information depending on its speaking style. As the phonetic information is basically only the vowel “e”, and eventually a consonant /h/ preceding the vowel, the different speaking styles are likely produced by mainly changing prosodic and voice quality features. Possible paralinguistic information (speech acts, attitudes or emotions) transmitted by varying the speaking styles of the utterance “e” is listed below.

affirmation (***aff***), agreement, understanding or consent (***agr***), backchannel (make agreeable responses) (***backch***), asking for a repetition (***askrep***), embarrassment (***emb***), dissatisfaction (***dissat***), surprise, amazed or astonished (***surp***), unexpected (***unexp***), blame or criticize (***blm***), disgust or dislike (***disg***), admiration or be impressed (***adm***), envy (***env***), suspicion (***susp***), thinking or filler (***thk***), sympathy (***symp***).

As the items of the list are difficult to be clearly separated in terms of intentions, attitudes, or emotions, the term "**speech act**" is used in this paper to refer to all items of the list.

Here, speech data is recorded in order to get a balance in terms of the paralinguistic information carried by the utterance "e". For that purpose, sentences are elaborated in such a way to induce the subject to produce a specific speech act. Two sentences are elaborated for each speech act item.

The sentences are first read by one native speaker. Then, subjects are asked to produce a target speech act, i.e., utter in a way to express a determined speech act, through the utterance "e", after listening to each sentence (pre-recorded "inducing" utterance). Some short sentences are also elaborated to be spoken after the utterance "e", in order to get a reaction as natural as possible, and a pause is requested between the utterance "e" and the following short utterance. Also, the utterance "he" (with the aspirated consonant /h/ before the vowel /e/) is allowed to be spoken, if the speaker judges that "e" is not appropriate for expressing some speech act.

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 speech act list, speakers are also asked to utter both "ee" and "hee" in a pressed voice quality, which frequently occurs in natural expressive speech [6], but was found more difficult to naturally occur in an acted scenario. For complementing the data above in terms of non-modal voice qualities, another speech data is also prepared, by selecting "e" utterances from a natural conversational database (JST/CREST ESP Project) of one female speaker (FAN) in her 30s, recorded during a long period of about 3 years. The criteria for selection from this natural speech data will be explained in Section 4.2.4. All the utterances "e" are manually segmented for subsequent analysis and evaluation.

3. Perceptual data of speech acts

Perceptual experiments are conducted on the "e" utterance data, to verify if the intended (induced) speech act can be correctly recognized when listening only to the utterance "e", i.e., in a context-free situation.

4 subjects are asked to choose one or multiple items, from the speech act list, that could be expressed by each of the 208 stimuli (segmented "e" utterances). A stimulus is counted as a perceived speech act, when 3 or more subjects perceive the same speech act.

The results below show how well the intended speech acts are correctly perceived by listening only to the utterance "e" (number of stimuli correctly perceived / total number of stimuli of an intended speech act).

- correctly perceived: *aff* (12/12), *agr* (9/9), *backch* (8/10), *askrep* (11/12), *adm* (10/12), *surp* (10/12), *th* (8/10)
- reasonably perceived: *disg* (8/12), *dissat* (7/12).
- poorly perceived: *emb* (2/12), *symp* (2/12), *unexp* (4/12), *blm* (5/12), *env* (5/12).

The mismatches or ambiguities between intended and perceived speech acts are also analyzed. Among the items with poorly perceived speech act, most of ***unexp*** is perceived as ***surp***, while most of ***emb*** is perceived as ***thk*** or ***dissat***. Confusion is also found between samples of ***blm, disg, dissat*** and ***susp***. Further, the items ***symp*** and ***env*** are perceived as several different speech acts, while stimuli of several different intended speech acts is perceived as ***dissat***. This implies that an automatic discrimination of these categories will probably be difficult based only on the speaking style of the utterance "e", i.e., in a context-free situation. Also as a result of the multiple choices for each stimulus, groups of speech act items can be roughly distinguished as shown in Fig. 2, where some confusion is found between the items within each group, but not so much confusion between the items of different groups.

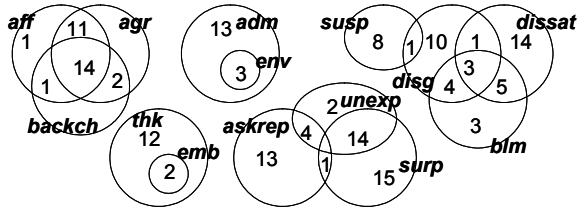

Figure 2: *Number of utterances perceived for each speech act items, and resulting grouping*.

The effects of context dependency for the perception of speech acts will not be discussed in this paper. Rather, the 157 stimuli (shown in Fig. 2) where perception agreement is obtained among the subjects will be used in the subsequent acoustic analysis, to show how much information can be extracted in a context-free situation, i.e., based only prosodic and voice quality features of "e" utterances.

4. Evaluation of acoustic parameters

In this section, we describe acoustic parameters that potentially represent the perception of features related with the different speaking styles, and evaluate them in the speech act discrimination.

4.1. Acoustic parameters related with prosodic features

In [7], a set of parameters was proposed for describing the intonation patterns of phrase finals (phrase final syllables), based on F0 and duration information. Here, we use a similar set of parameters, for the monosyllabic "e" 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 [7]. All F0 values are converted to the musical (log) scale before any subsequent processing.

The (monosyllabic) utterance is broken in two segments of equal length, and representative F0 values are extracted for each segment. In [7] several candidates for the representative F0 values were tested, and 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 extrapolated 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 [7].

Fig. 3 shows the distributions of *F0move* and duration for each speech act groups. From the distributions of the parameters shown in the figure, we can observe that groups of speech acts expressing some functions can be discriminated by using prosodic features: the group ***aff/agr/backch*** (positive functions) shows falling tones (negative *F0move*) and not long

durations around 300 ms; flat tones (*F0move* around zero) are found in ***thk/emb*** (filler-like functions); ***askrep*** (ask for repetition) shows short and rising tones, while ***susp*** (suspicion) shows longer and deeper rising tones. However, a large overlap is obtained in rising tones (positive *F0move*) between ***askrep*** and ***surp/unexp*** for short durations, and between ***adm/env, blm/disg***, ***dissat*** and ***surp/unexp*** for longer durations, indicating that the only use of the proposed prosodic parameters are not enough to discriminate between these speech act items.

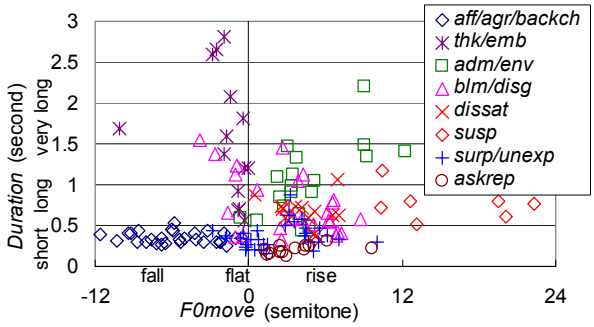

Figure 3: *Distributions of the prosodic parameters for each perceived speech act group.*

4.2. Acoustic parameters related with voice quality features

In this section, the use of parameters related with voice quality (non-modal phonations) is evaluated for a better discrimination between speech act items that can not be discriminated by the only use of prosodic features.

4.2.1. Detection of vocal fry (creaky) segments

Here, we use an algorithm recently proposed in [8] for detection of vocal fry segments. The algorithm searches for power peaks in a "very short-term" power contour (obtained by using 5 ms frame length each 2.5 ms), which reflects the pulse properties in very low fundamental frequencies, characteristic of vocal fry signals, and then check 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 algorithm can be found in [8].

4.2.2. Detection of segments with double-periodicity or aperiodicity

Both vocal fry (creaky) and harsh voices are characterized by irregularities in the periodicity of the vocal fold vibrations. These irregularities can be of double-periodic or aperiodic nature. Here, we use parameters proposed in [9] for detection of double-periodicity and aperiodicity, firstly proposed for creaky voice detection, but that could also be used to detect aperiodicities caused by harsh voices. The idea is to consider as harsh, the segments detected here as aperiodic or double-periodic, and not detected as vocal fry in 4.1.1.

The parameters are based on the relations between the first two peaks of the normalized autocorrelation function of the vocal tract inverse-filtered pre-emphasized speech signal, which provide information about the periodicity of the speech signal. One of the parameters, called ***NACR*** (Normalized AutoCorrelation Ratio), is the ratio of the normalized autocorrelation values of the first two peaks in the autocorrelation function. The second parameter, ***TLR*** (Time-Lag Ratio), is the ratio of the autocorrelation lags of these two peaks, multiplied by 2. *NACR* values larger than 1, or *TLR* values different from 1, indicate possibility of double-periodicity or aperiodicity. More details about the analysis and evaluation of these parameters can be found in [9].

4.2.3. Detection of segments with aspiration noise

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 [1], a categorical classification of voices in whispery or breathy is difficult in both acoustic and perceptual spaces [11]. Further, aspiration noise is also often perceived in harsh voices, which is called harsh whispery voice in [1]. So, in the present work, rather than classify the voice qualities, we use a degree of aspiration noise as indicative of such voice qualities.

The aspiration noise detection algorithm is based on the proposed in [10]. The algorithm depends basically on two parameters. 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 [10]. Here, aspiration noise is detected when *F1F3syn* is smaller than 0.4 and *A1-A3* is smaller than 25 dB.

4.2.4. Relationship between voice quality parameters and perceived speech acts

The detection algorithms introduced in the previous sections provide information about voice quality in frame level or in segmental level. The following parameters are then proposed for giving voice quality information in utterance level.

- Vocal fry rate (*VFR*): Proportion in duration of vocal fry (creaky) segments.
- Aperiodicity rate (*APR*): Proportion in duration of aperiodicity or double-periodicity (excluding the ones caused by vocal fry segments). This measure is expected to reflect the aperiodicities caused by harshness.
- Aspiration noise rate (*ANR*): Proportion in duration of aspiration noise segments.

Preliminary experiments in detection of each perceived voice quality indicated a threshold of 0.1 as being reasonable for detecting the above parameters. Thus, *VF* label is attributed if *VFR* > 0.1, *AP*, if *APR* > 0.1, *AN*, if *ANR* > 0.1,

and *M* (modal), otherwise.

Table 1 shows the distributions of detected and perceived voice qualities for the stimuli of each perceived speech act group. The perceived voice quality data is based on the annotation of voice quality by a subject with experience in voice quality (the author itself). Samples can be heard in the following link: <http://www.irc.atr.jp/~carlos/voicequality>.

Table 1: Number of utterances of detected (and perceived) voice qualities, for each perceived speech act group.

	VF	*AP*	*AN*	*M*
aff/agr/backch	1		4 (6)	24 (23)
thk/emb	2 (3)		1 (2)	11 (9)
adm/env	3 (4)		2 (2)	11 (10)
askrep			1 (1)	12 (12)
surp/unexp		6 (10)	12 (12)	18 (14)
susp		2 (3)	5 (5)	2 (1)
blm/disg/dissat	4 (4)	3 (9)	6 (5)	8 (6)
dissat			2 (2)	12 (12)

Results in table 1 show tendencies of non-modal voice qualities (*VF*, *AP* and *AN*) appearing in speech act groups expressing stronger emotion or attitude (***surp/unexp, susp, blm/disg/dissat, adm/env***). However the number of utterances of non-modal voice qualities is much smaller than in modal voice quality (*M*). Thus, in order to complement the above data in terms of voice quality, "e" utterances were additionally selected from a natural conversational database. The selection was realized by searching "e"/"he" from the text information, and selecting the ones whose acoustic detection matched with the perception of non-modal voice qualities, resulting in 60 more utterances (15 *VF*, 15 *AP*, and 30 *AN*). Figure 4 shows the distributions of each detected voice qualities.

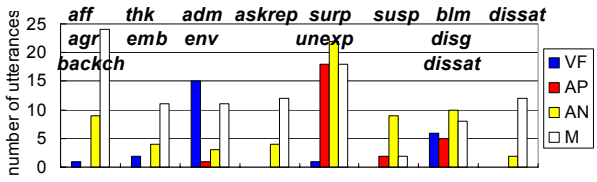

Figure 4: *Distributions of the detected voice qualities for each perceived speech act group.*

From the distributions of *AN* in Fig. 4, a strong presence of aspiration noise is clear in speech acts expressing strong emotions and attitudes like ***surp/unexp, susp, blm/disg/dissat***, relative to other voice qualities. The presence of aspiration noise in ***aff/agr/backch*** is probably related with politeness [5]. The distributions of *AP* show predominance in ***surp/unexp*** and ***blm/disg/dissat***. Finally, the distributions of *VF* show strong presence of vocal fry in ***adm/env*** and ***blm/disg/dissat***. Some samples of ***thk/emb*** and ***aff/agr/backch*** were detected as *VF*, but these samples were lax creaky in contrast with the pressed ones in ***adm*** and ***disg***. So, additional acoustic parameters are necessary for discriminating between pressed and lax voices. Further, most samples uttered by "he" (instead of "ee") were perceived as ***adm***, indicating that phonetic information is also important for speech act discrimination.

Regarding the automatic detection of voice quality, comparing the numbers inside and outside the parenthesis in Table 1 for *VF*, *AP* and *AN* columns, it can be observed that the tentative of detecting harsh voices using the aperiodicity parameters of 4.2.2 resulted in a correct detection of part of the utterances present in ***surp*** and ***blm***/***disg***. Therefore, improvements are necessary for a better acoustic characterization of harsh voices.

5. Conclusions

Perceptual experiments on speech acts and speaking styles of the utterance "e" indicated that prosodic features are effective for identifying some groups of speech acts with specific functions (affirmation/agreement/backchannel, ask for a repetition, fillers), while voice quality features are more effective for identifying speech act items expressing some emotion or attitude (surprise, disgust, suspicion, dissatisfaction, admiration).

The goal of the present work is to automatically detect the different speech act items or groups. Although the acoustic analysis showed that part of the speech act groups can be correctly discriminated using the proposed acoustic parameters, improvements are still necessary mainly in the parameters related with voice quality features.

Future works are to improve the voice quality parameters, and evaluate automatic detection of speech act groups, using the whole set of prosodic and voice quality parameters.

6. Acknowledgements

This research is supported by the Ministry of Internal Affairs and Communications. We thank Ken-Ichi Sakakibara (NTT) and Parham Mokhtari (ATR) for advice and motivating discussions. We also thank ATR/HIS Labs for lending the sound-proof booth for recordings.

7. References

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

[2] Erickson, D., 2005. Expressive speech: production, perception and application to speech synthesis. *Acoust. Sci. & Tech*., Vol. 26 (4), 317-325.

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

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

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

[6] Sadanobu, T., 2004. A Natural History of Japanese Pressed Voice, *J. of Phonetic Society of Japan*, Vol. 8 (1): 29-44.

[7] Ishi, C.T.; Mokhtari, P.; Campbell, N., 2003. Perceptually-related acoustic-prosodic features of phrase finals in spontaneous speech, Proc. *Eurospeech 2003*, 405-408.

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

[9] Ishi, C.T., 2004. Analysis of autocorrelation-based parameters for creaky voice detection, Proc. *Speech Prosody 2004*, 643-646.

[10] Ishi, C.T., 2004. A new acoustic measure for aspiration noise detection, Proc. *ICSLP 2004*, Vol. II, 941-944.

[11] Kreiman, J.; Gerratt, B., 2000. Measuring Vocal Quality, In *Voice Quality Measurement*, Singular Thomson Learning, 73-102.