Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2010, Article ID 528193, 12 pages
doi:10.1155/2010/528193

Research Article

Analysis of the Roles and the Dynamics of Breathy and Whispery Voice Qualities in Dialogue Speech

Carlos Toshinori Ishi, Hiroshi Ishiguro, and Norihiro Hagita

Intelligent Robotics and Communication Laboratories, ATR, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

Correspondence should be addressed to Carlos Toshinori Ishi, carlos@atr.jp

Received 1 June 2009; Revised 11 September 2009; Accepted 25 November 2009

Academic Editor: Vijay Parsa

Breathy and whispery voices are nonmodal phonations produced by an air escape through the glottis and may carry important linguistic or paralinguistic information (intentions, attitudes, and emotions), depending on the language. Analyses on spontaneous dialogue speech utterances of several speakers show that breathy and whispery voices are related with the expression of a variety of emotion- or attitude-related paralinguistic information. Potential acoustic parameters for characterizing breathy/whispery segments are introduced and used to describe the dynamics of breathiness along the utterances in different paralinguistic items.

1. Introduction

Besides the linguistic information, the understanding of paralinguistic information (intentions, attitudes, and emotions conveyed by nonverbal elements of communication), is also important in spoken dialogue systems. Although prosodic features (features expressing the intonation, stress, and rhythm of utterances), like fundamental frequency (F0), power, and duration, have important roles in carrying paralinguistic information, analyses of natural conversational speech data have shown that variations in voice qualities (changes in the quality of the voice due to non-modal phonations, such as breathy, whispery, creaky and harsh voices [1]) are commonly observed, mainly in expressive speech utterances [2].

In a previous work [3], we have proposed a framework for extraction of paralinguistic information, considering both intonation-related prosodic features and voice quality features, as shown in Figure 1. However, evaluations have shown that a one-to-one mapping between paralinguistic information items (including intentions, attitudes and emotions) and the prosodic and voice quality features is difficult.

In the present work, we focus on breathy and whispery voice qualities, which are characterized by an auditory impression of turbulent noise, caused by an air escape through the glottis, and analyze their communication roles (i.e., the variations in paralinguistic information) in spontaneous dialogue speech, for several speakers. Note that fricative consonants are also characterized by turbulent noise, but it is produced by a constriction in the vocal tract, while turbulence in breathy and whispery voices is produced by a constriction at the glottis.

Breathy and whispery voices have been reported in the literature to carry important linguistic and paralinguistic information, depending on the language. For example, a phonemic contrast between breathy and modal voicing among vowels is particularly common in many minor languages [4]. In [5, 6], relationships between different phonation types and paralinguistic information like emotions and attitudes are reported. In [5], whispery voice was found in "fear" while breathy voice was found in "sad" voice. Correlations between synthesized breathy and whispery voices and the perception of relaxed/stressed, sad/happy, and intimate/formal are reported in [6], for English. Breathiness is also reported to appear in the expression of disappointment, for Japanese [7, 8]. In Japanese spontaneous speech, possible use of breathiness for expressing manner or politeness is reported [9]. However, none of these works have analyzed the dynamics, that is,

how the breathy and whispery components change along the utterance.

Regarding the terminologies, "breathy" and "whispery" phonations can distinctly be defined from a physiological viewpoint [1]. Figure 2 shows the laryngeal tensions involved in breathy and whispery phonations, compared to the modal (normal) phonation. In whispery phonation, the glottal constriction results from a triangular opening of the arytenoid cartilages of the glottis, while the vocal folds can vibrate independently. This triangular opening results from a weak adductive tension and a moderate-to-high medial compression. If the ligamental part of the vocal folds vibrates, "whispery voice" is produced, while if it does not vibrate, (unvoiced) "whisper" is produced. In breathy phonation, all laryngeal muscles are relaxed, so that the glottal constriction results from an incomplete closure of the vocal folds during vibration.

Breathiness is generally treated as a continuum that is difficult to separate into "breathy" and "modal" whether sorting is based on perceived quality, acoustics, or the underlying glottal configuration [10]. The transition from "breathy" to "whispery" seems also to be part of an auditory continuum [1]. Although breathy and whispery voices have distinct definitions in terms of the phonation settings, they are often confused, probably because they are similarly characterized by the auditory impression of turbulent noise (aspiration noise).

In the present work, considering that our voice quality data is based on auditory impression, we use the terms "breathy" and "breathiness" in a broad sense, indicating all utterances where turbulent noise is audibly perceived in the vowel segments. However, the term "whispery" is also used in the paper, when the auditory impression of the turbulent noise is closer to whisper, rather than to normal phonation.

The rest of the paper is organized as follows. In Section 2, the production and acoustic properties of breathy/whispery voices are explained, and some acoustic parameters for their characterization are introduced. The spontaneous speech data and the annotation data are described in Section 3. In Section 4, relationship between the introduced acoustic parameters and the perceptual data is analyzed. In Section 5, the speaking styles are analyzed for each group of paralinguistic information, considering the dynamics of breathiness, intonation, and linguistic information. Discussions are presented in Section 6 and the paper is concluded in Section 7.

2. Acoustic Features Characterizing Breathy and Whispery Voices

In this section, we introduce acoustic features which are intended to characterize breathy and whispery voices. Although the main focus of the present work is to investigate the paralinguistic roles of breathy and whispery voices (via perceptual tasks, as will be described in Section 3), we discuss the potentiality of some acoustic features for quantifying the degree or intensity of breathiness (in Section 4), and for describing the temporal patterns of breathiness along the utterances (in Section 5).

Both breathy and whispery phonations are characterized by an air escape through the glottis, which increases the noisy (turbulent, nonharmonic) components in the frequency bands around the third formant [11, 12]. For relatively small average glottal openings, as in modal phonation, the noisy component is much lower in amplitude than the periodic component; while for relatively large glottal openings, voicing does not occur and only noise (aspiration) is generated. For intermediate average glottal openings, as in breathy phonation, voicing may continue, but two changes occur: one is a high spectral damping in higher harmonics; the other is an increase in the amplitude of the turbulent noise because of the increased flow. Figure 3 shows a scheme of how periodic and noisy spectral components change in modal and breathy voices [11].

There are a number of acoustic parameters reported in the literature, which try to characterize the acoustic properties of breathy/whispery voices. For example, *H1-A3* (difference between the amplitudes of the first harmonic and the third formant) [13] and *NAQ* (normalized amplitude quotient of the glottal waveform and its derivative waveform) [14] are reported to be correlated with breathiness. However, these parameters can only characterize the spectral slope properties of breathy voice, regardless of the presence of aspiration noise components characteristic of breathiness. Other parameters such as *HNR* (harmonics-to-noise ratio), *GNE* (glottal-to-noise excitation ratio) [15], and $f_{\text{aperiodic}}$ (boundary frequency between harmonic and aperiodic components) [16] reflect the effects of the aspiration noise components. However, *HNR* and $f_{\text{aperiodic}}$ depend on harmonicity information, being less reliable in segments where pitch changes or where the glottal pulses are irregular. The *GNE* measure is more robust to such segments, but depends on vocal tract inverse filtering to get estimates of the glottal excitation, being less reliable for high-pitched voices.

In our previous work, we proposed a new measure for aspiration noise characterization called *F1F3syn* (synchronization of the amplitude envelopes of the first and third formant frequency bands) [17]. *F1F3syn* is similar to *GNE*, but one of the differences is not doing inverse filtering to avoid its problems in high-pitched voices. *F1F3syn* will be described in detail in Section 2.1.

Finally, although all above parameters provide information about the presence of breathiness, they do not provide information about its intensity. Thus, in the present work, we make use of the parameter *F1F3syn*, and estimate a breathiness power measure, as will be described in Section 2.2.

2.1. The F1F3syn Measure. The *F1F3syn* measure is based on the idea that the amplitude envelopes of the signal filtered around the first and the third formant frequencies (F1 and F3) are synchronized in modal segments (where the vocal tract responses are synchronized with impulse-like excitation) and unsynchronized in breathy segments (where a turbulent noisy excitation is prominent around the third formant).

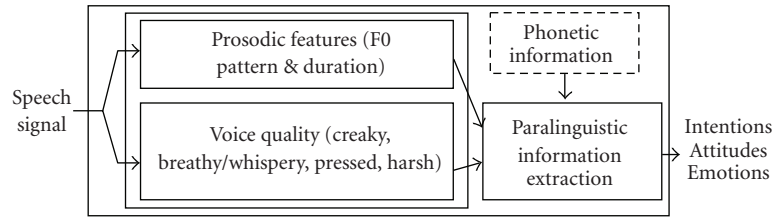


Figure 1: A framework for paralinguistic information extraction considering intonation-related prosodic features and voice quality features.

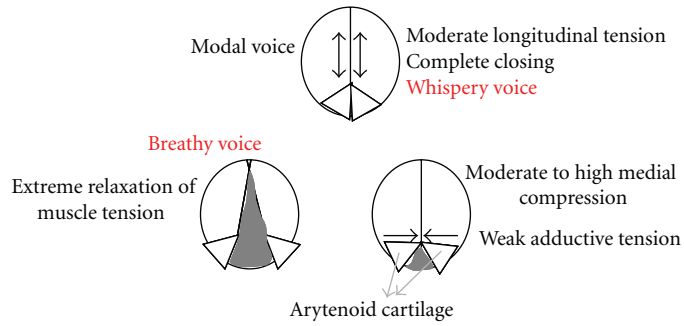


Figure 2: Phonation settings for modal, breathy, and whispery voices.

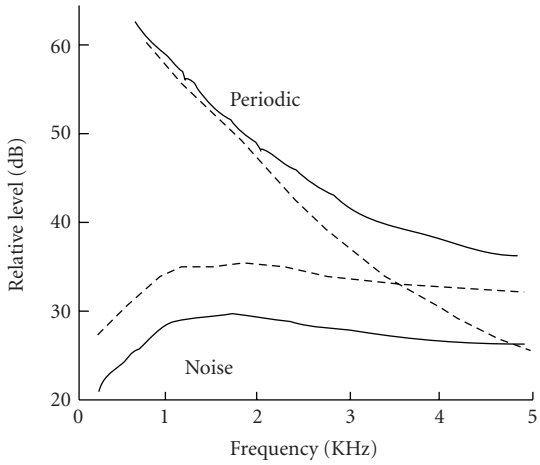


Figure 3: Periodic and noisy components in modal (full lines) and breathy voices (dashed lines) [11].

Figure 4 shows a block diagram of the proposed measure. First, bandpass filtering is realized by taking the Fourier transform of the input signal in a 32 ms frame length, and windowing each frequency band (in the frequency domain). In order to avoid automatic formant extraction, we set a fixed and broad range for each frequency band. For the F3 band, we set a range of 1500~4000 Hz, which is likely to contain F3 (the third formant) for both male and female speakers. For the low frequency band, we set a range of 100~1500 Hz, since periodicity is less affected by aspiration noise in this region [11]. Although a band of 1000 Hz is suggested in [5], we decided to use a higher value of 1500 Hz. This is to guarantee that at least two harmonics will be present in the frequency range, avoiding the glottal excitation pulsing to be missed. Note that if the frequency range is limited to 1000 Hz, only one harmonic (pure sinusoid) would be present for signals with very high fundamental frequencies (above 500 Hz), such that the amplitude envelope would be a constant, and information about the excitation pulsing would be missed. We refer to the low frequency band as "F1 band", since this frequency range is likely to contain the first formant for both male and female speakers. The bandpass filtered signals are referred to as *F1wave* and *F3wave*.

Next, the amplitude envelopes of the bandpass filtered signals are estimated for each frequency band. First, the Hilbert envelopes [18] of each frequency band are estimated, for providing instantaneous amplitudes of the signals. The Hilbert envelope is defined as the magnitude of the analytic signal, whose real part is the signal and the imaginary part is the Hilbert transform of the signal (which is the signal with its phase shifted by 90 degrees). In practice, the Hilbert transform is realized in the frequency domain, by doubling the positive frequency components, and putting zeros in the negative frequency components. The amplitude envelopes (*F1env* and *F3env*) are then obtained by smoothing the Hilbert envelopes of each band by using a Hann window of 1 ms length. Such a smoothing process is necessary because the F1 and F3 bands have different bandwidths.

Then, the amplitude envelopes of F1 and F3 bands are cross-correlated to obtain an index of synchronization (*F1F3syn*) between these signals. The region for the cross-correlation calculation was set to the 25 ms center region of the frame, where the signals are more stable. If the index of synchronization is low (i.e., the signals are uncorrelated), there is a high probability that the noisy components of the F3 band were produced independently of the glottal excitation pulses. Therefore, the input signal is likely to contain breathiness.

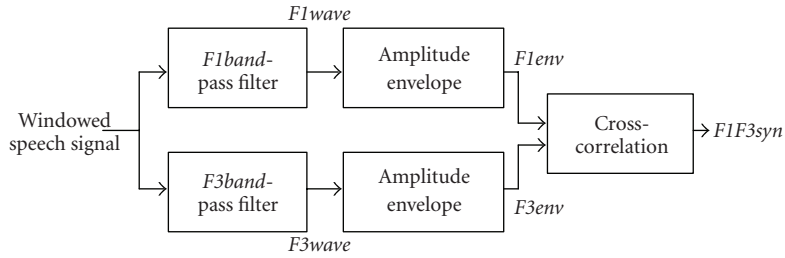


Figure 4: Block diagram of the *F1F3syn* computation.

Figure 5 shows examples of the bandpass filtered signals and the respective amplitude envelope signals for several voice qualities of the vowel /a/, extracted from a female voice of our speech database. The amplitude envelope signals are normalized in the figure to allow a more suitable visual comparison. Figure 5(a) shows the signals for modal voice. We can observe synchronized amplitude envelopes for *F1wave* and *F3wave*, and also in the estimated envelopes *F1env* and *F3env*. The index of synchronization *F1F3syn* in this case is 0.75. Figure 5(b) shows an example of breathy voice. A clear periodicity can be observed in the *F1wave*, but no clear regularities can be observed in *F3wave*. The amplitude envelopes *F1env* and *F3env* also reflect different shapes, and the *F1F3syn* value is low (0.05). Figure 5(c) shows an example of creaky voice (or vocal fry), which is a phonation type characterized by impulse-like glottal excitation usually accompanied by fundamental frequencies lower than the normal phonation, and commonly associated with lower aerodynamic pressures at the glottis. As creaky phonation is often realized in very low fundamental frequencies, it may happen that only one glottal pulse appears in the analysis frame, as the example in Figure 5(c). In this case, even though there is no periodicity in the analysis frame, we can observe similar shapes for the envelopes *F1env* and *F3env* (due to a synchronized response to the excitation pulse), and consequently a high *F1F3syn* of 0.87, showing that *F1F3syn* is also able to discriminate breathy from creaky segments.

2.2. Breathiness Power. Although the *F1F3syn* measure described in the previous subsection gives an estimation of presence or absence of breathiness, it does not provide information about the intensity (power) of the breathiness. Further, there are two problems in using *F1F3syn* for identifying breathy/whispery voices. One is that if the power of the F3 band is too low, the noise components (i.e., the breathiness) could not be perceived, so that it would not make sense to take a synchronization measure of frequency components which are not audible. The other problem is that the turbulent noise could be due to fricative consonants, like /s/ and /sh/, instead of breathiness.

To account for these problems, we defined a *breathiness power*, as the power of the F3 band (*F3 power*), constrained by a high *F1F3syn* and by a high *fricative power*.

The power of the F3 band is estimated as the RMS value of *F3wave* in dB. The *fricative power* is estimated as the RMS value of the signal filtered at the range of 4000 ~ 8000 Hz. In practice, the power values of these bandpass filtered signals are obtained by averaging the squared magnitude spectral components in the specified frequency band, and then converting them to dB. *Breathiness power* is then defined as

$$\text{Breathiness power} = \begin{cases} F3 \text{ power} & \text{if } A \text{ and } B \\ 0, & \text{otherwise,} \end{cases}$$

where condition A is

$$\text{fricative power} - F3 \text{ power} < \text{fricative threshold}, \tag{1}$$

and condition B is

$$F1F3syn < \text{synch threshold}.$$

In this way, *breathiness power* is expected to have high values in breathy/whispery segments, low values in segments where the noisy components cannot be clearly perceived, and null values in nonbreathy voiced and fricative segments.

3. Speech and Annotation Data

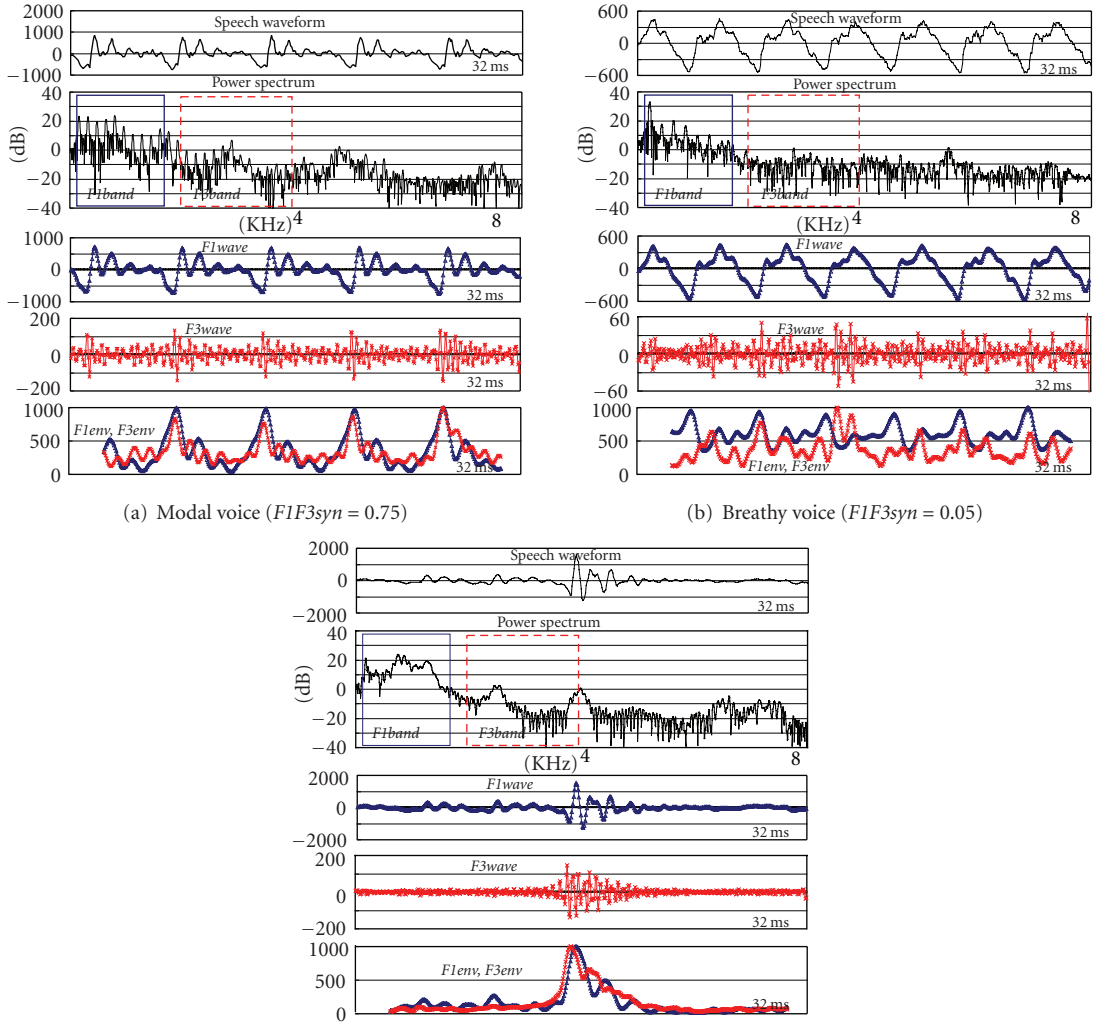
Two databases of Japanese spontaneous speech were analyzed in the present work. One is the JST/CREST ESP expressive speech database [19]. Data of eight speakers (six female and two male speakers) were selected for analysis. This database can be classified in two types.

(i) FAN, FYM, FSM, FYS (female, 30 seconds): natural daily conversations (including telephone calls) between family members, friends, and nonfamiliar people (hospital, companies). The length of each dialogue file varies from 10 to 30 minutes.

(ii) JFA (female, 40 seconds), JFB (female, 30 seconds), JMA (male, 20 seconds), JMB (male, 30 seconds): free dialogue conversations (by telephone) between subjects who were not familiar with each other. Each dialogue file has approximately 30 minutes.

Two to four dialogues were randomly selected from the database of each speaker, corresponding to one to two hours of dialogue data for each speaker.

The other database is the CSJ (Corpus of Spontaneous Japanese) speech corpus [20]. This corpus is constituted



(a) Modal voice (*F1F3syn* = 0.75)

(b) Breathy voice (*F1F3syn* = 0.05)

(c) Creaky voice (*F1F3syn* = 0.87)

FIGURE 5: Speech waveform, power spectrum, filtered signal in F1 band (F1wave), filtered signal in F3 band (F3wave), and amplitude envelopes of F1 and F3 bands (F1env and F3env), for representative segments of: (a) modal voice, (b) breathy voice, and (c) creaky voice.

by monologue and dialogue speech data. For the present work, dialogue data was analyzed. The dialogues are between speakers who are familiar and not familiar with each other. Each dialogue file has approximately 10 minutes.

(i) D01 (16 dialogues; total 3.2 hours): interview after simulated public speaking.

(ii) D02 (16 dialogues; total 3.1 hours): task-oriented dialogues.

(iii) D03 (16 dialogues; total 3.6 hours): free dialogue conversations.

(iv) D04 (10 dialogues; total 2.1 hours): interview after academic presentations.

Speakers are male and female aging from 20 seconds to 50 seconds. The interviewers or conversation partners of each dialogue are two female speakers, one in her 20 seconds and the other in her 30 seconds.

We used the utterance units provided by each database, which may contain one or more phrases. The dialogues selected from the two databases resulted in a total of 21819 utterances.

3.1. Annotation of Perceived Breathiness. Two subjects (with no experience in voice quality annotation) listened to the utterances, and identified the portions where breathy/whispery voices (hereinafter, br/wh) are perceived. Note that the term "br/wh" is used to indicate all speech segments where turbulent noise is perceived in vowel segments, including breathy voice, whispery voice, (unvoiced) whisper, and (unvoiced) aspirated sounds. During about a period of one week, the annotations of the two subjects were supervised by an expert in voice quality classification (the first author). The main problems arisen during supervision were that subjects had difficulties in discriminating breathy/whispery voices from the aspirated consonant /h/ and from lengthened fricatives /s/ and /sh/.

One of the subjects identified breathiness in 1584 utterances, while the other subject identified it in 1752 utterances. The 1134 utterances, where both subjects identified the presence of breathiness, were used for subsequent analysis.

3.2. Segmentation and Annotation of Perceived Voice Qualities. For the utterances where breathiness was perceived (excluding the laughing speech utterances), a more detailed segmentation and annotation of voice quality was conducted for evaluating the acoustic features introduced in Section 2. The segmentation was conducted by the first author, based on visual inspection of the spectrograms and on auditory impression. The segment categories are "br/wh voiced" (for breathy and whispery voiced segments), "br/wh" (for unvoiced whispered or aspirated segments), "br/wh?" (for segments with acoustic and auditory properties intermediate between breathy/whispery and other voice qualities), "modal" (for normal phonation in voiced segments), "fricative" (for fricative and affricative consonants), "aspirated consonant" (for /h/), "nasal", and "rough" (for rough quality segments, including vocal fry, creaky, and period-doubled segments).

3.3. Annotation of Paralinguistic Information. Paralinguistic information (PI) was annotated by two subjects (one with some experience in PI annotation, and the other without any experience), for the 1134 utterances where breathiness was perceived. In the present work, a previously prepared list of PI items (based on [3, 7–9]) was given to the subjects, but new items were allowed to be freely added, according to the subject's impression. The first set of PI items included the following 14 items: surprise, admiration, anger, fear, disgust, joy, sad, funny, dissatisfaction, suspicion, politeness, tiredness, disappointment, and confidential talking. The free annotation of PI items by the subjects resulted in an inclusion of 14 more new items: forced laugh, bitter laugh, excitement, emphasis, calling for attention, interest, gentleness, real feeling expression, sympathy, keenness, emotion quoting, diffidence, undecided, and talking-to-oneself.

Although there are many works related to emotion classification (such as [21] for European languages), we preferred to use an open set of response items, as described above, since the appearance of breathiness must not be restricted to emotions, but may also be related to other paralinguistic information, such as attitude-related items. Further, the relationship between voice quality and PI may change according to the language.

A preliminary comparison of the raw labels resulted in a matching of 38.9% between the labels of the two subjects. The labels were then revised by three subjects (the same two subjects who annotated the PI, plus another subject with experience in PI annotation) together, in order to match the items which could express close meanings. After the revision, the matching rate increased to 69.5%. Regarding the mismatches, different types of laugh were attributed by the two subjects in 9.0%, no labels could be attributed by one of the subjects in 3.8% of the utterances, while the rest are other types of mismatch. The distribution of the PI items where matching was obtained after revision by the three subjects are shown in Figure 6.

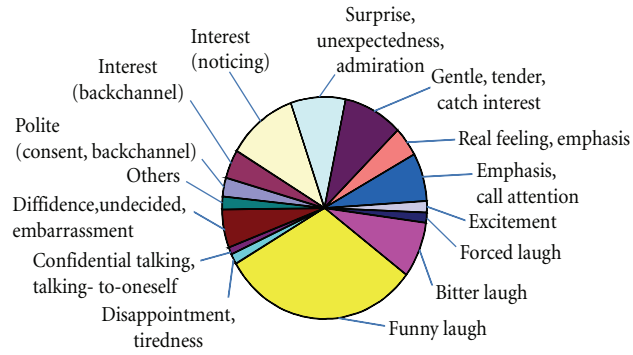


Figure 6: Distribution of the paralinguistic information items carried by breathy/whispery voices.

Speech samples of each PI item can be listened in the following homepage: <www.irc.atr.jp/~carlos/breathywhispery/>.

4. Relation between Perceived Breathiness and Acoustic Parameters

Figure 7 shows the distributions of the acoustic parameters *F1F3syn*, *fricative power − F3 power*, and *breathiness power*, for each segment type (described in Section 3.2). The distributions are normalized by the total length for each segment type.

The distributions of *F1F3syn* in Figure 7(a) show concentration in high values for "modal" and "rough" segments, concentration in low values for "br/wh" and "br/wh voiced", and intermediate values for "br/wh?", indicating consistency with the manual labels. A threshold around 0.5 seems to be reasonable for their discrimination. However, "fricative", "nasal" and "aspirated consonant" also show distributions similar with the breathy categories, indicating that the only use of *F1F3syn* is not enough for their discrimination. The distributions of *fricative power − F3 power* in Figure 7(b) show distinct distributions between "fricative" and other segment types. A threshold around 0 dB indicates reasonable separation of fricative from the other segments. Finally, the distributions of *breathiness power* in Figure 7(c) show a rough separation of "nasal" and "aspirated consonant" (having lower *breathiness power* values) from the breathy segments. It can also be noted that a high ratio of "modal", "rough" and "fricative" segments have null *breathiness power* values, thanks to the *F1F3syn* and *fricative power* constraints. These results show that *breathiness power* can potentially be used to quantify the intensity of perceived breathiness.

Figure 8 shows spectrograms, fundamental frequency (F0), *F1F3syn* and *breathiness power* contours for representative speech utterances, found in our spontaneous speech database, where br/wh segments (indicated by arrows) appear in the expression of several paralinguistic information items. These figures will be used for describing the speaking styles appearing in each paralinguistic items, later

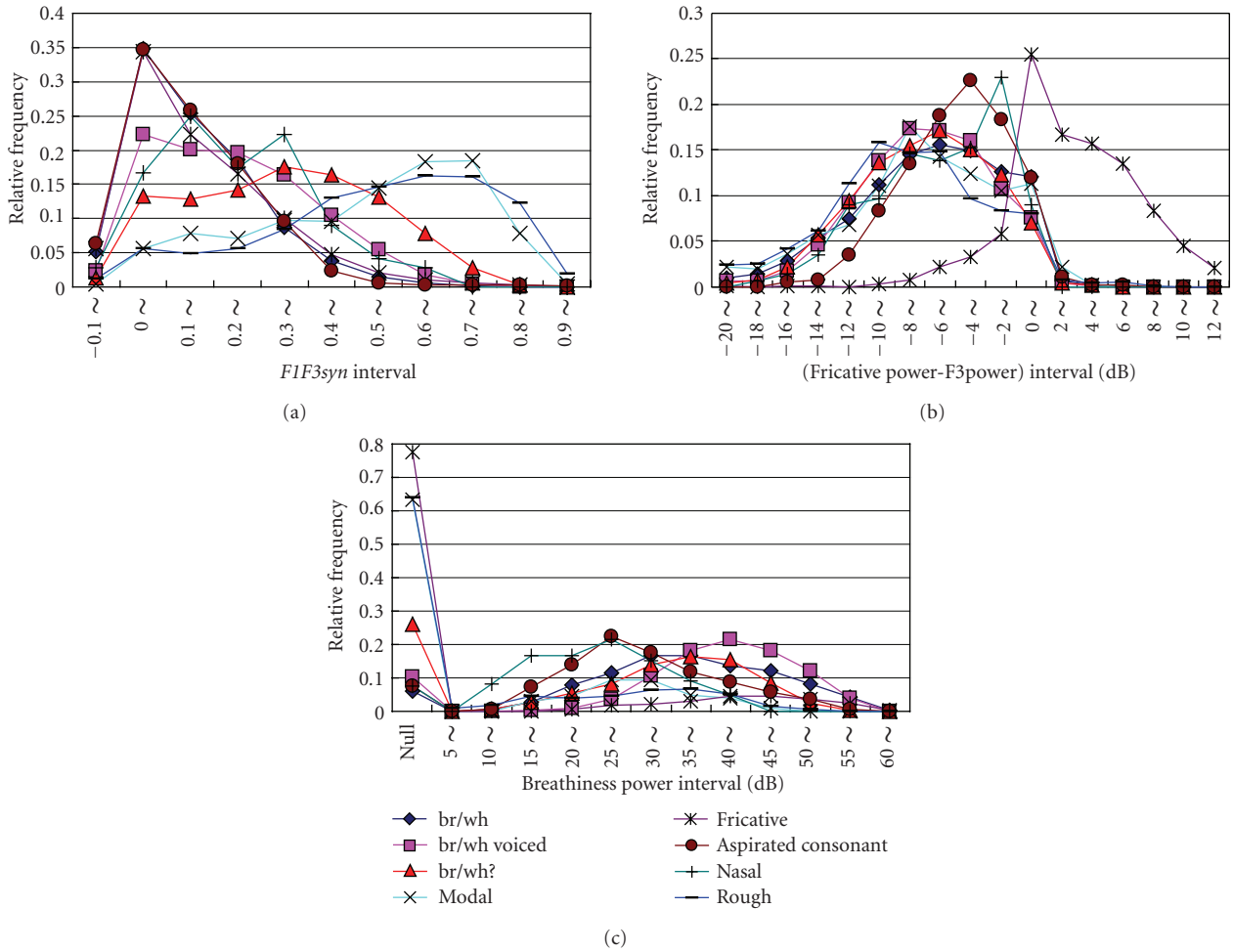


Figure 7: Normalized distributions of the acoustic parameters (a) *F1F3syn*, (b) *fricative power − F3 power*, and (c) *breathiness power*, for each segment type.

in Section 5. In the figures, the vowels perceived as br/wh are highlighted. It can be observed in the spectrograms that the harmonic components are present in the low frequency components around 0 to 1500 Hz (horizontal stripes), but they become weaker or absent in the range around 1500 to 4000 Hz, in the br/wh segments. In whispered and aspirated segments, where the vocal folds are not vibrating, harmonic components are also absent in the 0 to 1500 Hz frequency band. Note also that the segments with high *F1F3syn* values are zeroed in the *breathiness power* contours. In the present analysis, the *fricative threshold* was set to 0 dB, while the *synch threshold* was set to 0.5, based on the acoustic analyses above. The arrows in Figure 8 were manually put in the segments with high *breathiness power* values, in order to facilitate a visual comparison with the spectrograms and the highlighted vowels (perceived as br/wh).

5. Analysis of the Paralinguistic Information (PI) Carried by Breathy and Whispery Voices

In the following subsections, the PI items (the roles of the br/wh) are grouped, and the speaking styles considering intonation and temporal patterns of the br/wh segments within the utterances are analyzed. It is worth to remind that the PI items are language/culture dependent, and some of them may be specific for Japanese. However, similar methodology can be applied to analyze the speaking styles for any language.

5.1. Emphasis, Attention, Real Feeling Expression. The analyses of the present spontaneous speech database have shown that breathiness often appeared in emphasized (focused or prominent) word/phrases, and has the effect of calling/catching the attention of the listener. Figure 8(a) shows an example where breathiness appears in the emphasized word "*sootoo*" ("quite").

It is known that there are a number of ways for emphasizing a word/phrase while speaking. For example, one is raising the pitch in the focused word/phrase, another is increasing the power, and another is lengthening the word [22]. However, in some of the utterances annotated as "emphasis", the only use of breathiness, without a significant raise in pitch or power, was effective for expression of emphasis (e.g., Figure 8(a)).

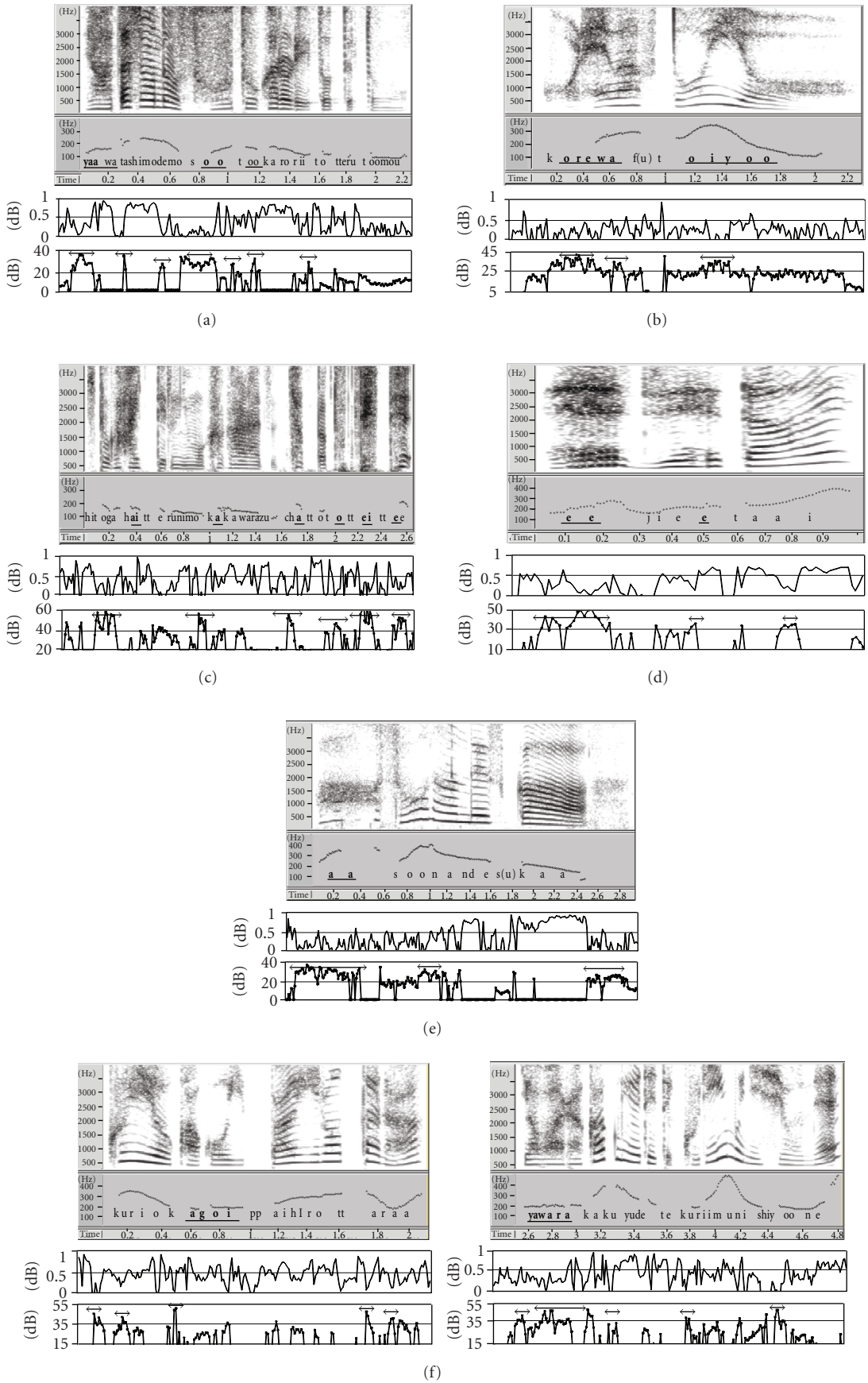


FIGURE 8: Continued.

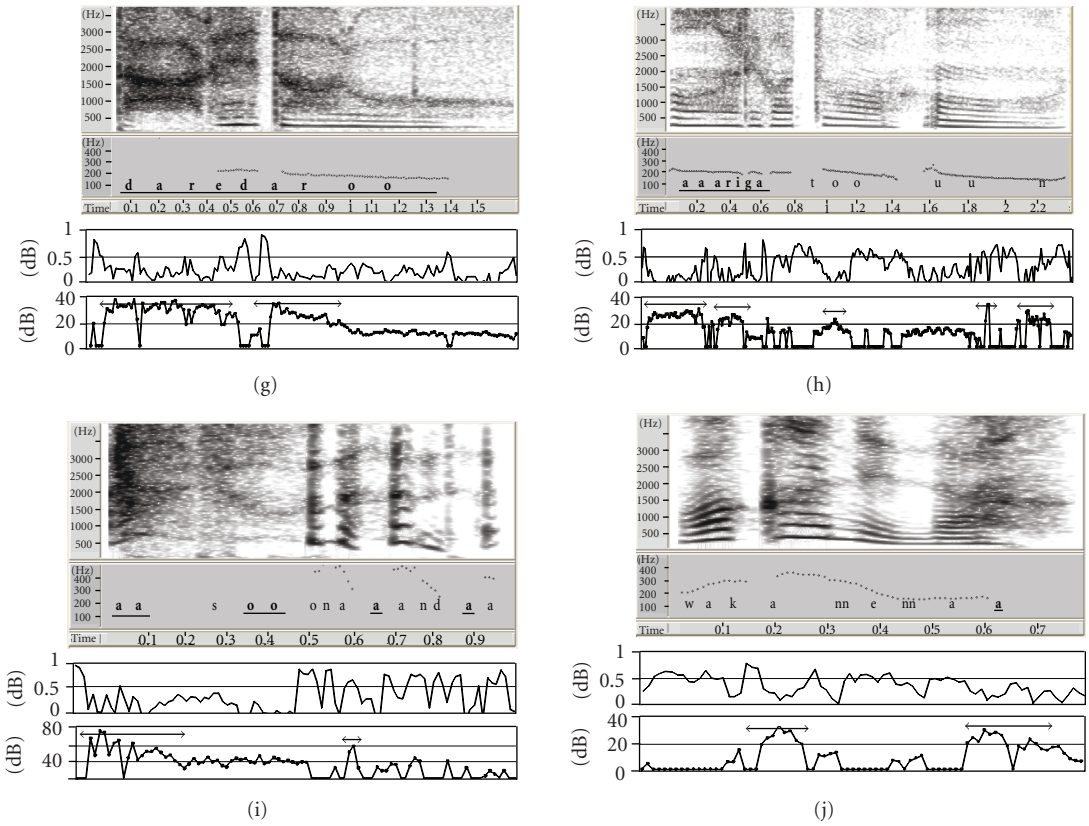


Figure 8: Examples of spectrograms, F0, *F1F3syn,* and *breathiness power* contours for spontaneous speech utterances including breathy/whispery segments (indicated by arrows), carrying several paralinguistic information: (a) emphasis/attention; (b) real feeling expression; (c) excitement; (d) surprise/unexpectedness; (e) surprise/admiration; (f) gentle/tender (story reading); (g) embarrassed/thinking/talking-to-oneself; (h) sighing speech; (i) funny laughing speech; (j) bitter laughing speech.

Regarding the speaking style, breathiness appeared frequently in the high pitch accented portions (80% of the "emphasis" tokens), and at the phrase beginnings (12% of the "emphasis" tokens).

Part of the utterances annotated as "emphasis" was also interpreted as if the speaker was expressing a "real feeling" (20% of the "emphasis" tokens). For example, the utterance "*kore wa futoi yo*" ("this is thick") accompanied by breathiness (Figure 8(b)) can be interpreted as "this is really thick" or "I really think this is thick". This usage of breathiness seems to have similar effects with pressed voices (which are characterized by a hyperclosure of the vocal folds), as reported in [23]. The effects of real feeling expressions seem to be stronger, as the speaking style approximates to whisper (absence of voicing).

5.2. Excitement. Along with "emphasis", some of the utterances were annotated as "excited". The br/wh segments rhythmically appeared within and across the utterances, when the speaker was speaking excitedly. Figure 8(c) shows an excited speech utterance containing br/wh segments. It can be observed that modal and br/wh segments occur alternately along the utterance. This speaking style was observed in 82% of the "excited" tokens, while in 10% of the "excited" tokens, only the first syllable of the utterance was breathy.

Changes in voice quality also happened when the speaker quoted an emotional (excited) speech. In the present data, br/wh voice appeared in quoted utterances with a speaking style similar to that in the (spontaneous) excited speech (all "quoted" tokens).

Further, br/wh voice in excited speech was observed in both positive (joy, cheerful) and negative (dissatisfied, irritated) emotional states.

5.3. Politeness and Interest in Backchannels. Politeness and interest were annotated in most of the interjections "*hai*", "*ee*" and "*un*" where breathiness was perceived. These interjections are commonly used as backchannels in dialogue when spoken by a short falling intonation. "*hai*" ("yes") is a more formal backchannel, "*un*" ("uhm") is a more casual backchannel, and "*ee*", is something in between.

Results showed that breathy "*hai*" and "*ee*" utterances are perceived as more formal/polite than their nonbreathy counterparts, while breathy "*un*" utterances express more interest to the interlocutor's talk, than their nonbreathy counterparts. Further, although "*un*" is often spoken as a nasalized schwa vowel in very casual situations, which could

express a sloppy manner, it was also found to express interest, when spoken with a breathy voice quality. Regarding the breathiness styles, 85% of "*hai*" tokens expressing politeness showed breathiness in the first half of the syllable (strong aspirations in /h/ and half of the vowel /a/), while in the remaining tokens, breathiness appeared in the whole utterances. In the case of "*un*", breathiness appeared in the whole utterance in 95% of the tokens.

The short interjection "*ah*", which usually expresses noticing, is commonly followed by backchannels "*soo*"/ "*soodesuka*" ("really?"), and is also often accompanied by breathiness. In such cases, breathy "*ah*" was also found to express more interest than the nonbreathy counterparts. Regarding the breathiness styles of "*ah*" tokens, an aspirated quality in the end portion of "*ah*" was found in 85% of the tokens, while in the remaining tokens, breathiness appeared in the whole "*ah*" utterance.

5.4. Surprise, Unexpectedness, Admiration. Surprise, unexpectedness and admiration may coexist, so that it is difficult to clearly separate them. However, the utterances could be roughly separated in two groups: surprise/unexpectedness (70% of the tokens; e.g., Figure 8(d)) and surprise/admiration (30% of the tokens; e.g., Figure 8(e)).

In surprise/unexpectedness, three types of speaking styles were predominant: br/wh voice along the whole word/phrase (40% of the tokens), aspiration at the end of the utterance (26% of the tokens), and a harsh voice quality (due to irregularity in the vocal fold vibrations) along with breathiness (20% of the tokens). The harsh voice quality tends to appear when the excitement level of the speaker increases. Figure 8(d) shows an example where the interjection "*ee*" is accompanied by a harsh whispery voice quality.

In surprise/admiration, br/wh voice along the whole word/phrase was predominant (80% of the tokens), with the breathiness approaching to a whispered quality. Note that the harmonic components are much weaker in the "*aa*" portion of Figure 8(e).

Almost all (95% of the) utterances annotated as surprise, admiration or unexpectedness, are accompanied by interjections or interjectional expressions like "*eeh*!", "*sugoi*!", "*hontoo*!", "*hee*!", "*haa*!", "*waa*!", "*aah*!", "*soonandesuka*!", "*soonanda*!", "*naruhodo*!", "*uso*!", which can equivalently be translated as "wow!", "really!?", "amazing!", "you are kidding!". Such linguistic information is thought to be important to discriminate them from the excited speech of Section 5.2.

5.5. Gentleness/Tenderness, Calling Attention. Breathiness accompanied by a soft voice quality appeared rhythmically along the utterances of the speaker FYM (mother), when she was reading stories to her child. This change in voice quality (from the speaker's normal speaking voice) is found to display expressivity and is thought to have the effect of calling/catching the attention of the listener, while expressing gentleness/tenderness. This speaking style could be related to a "code switching" where individuals modify their voices in the presence of young children.

Regarding the rhythm of breathiness within an utterance, spectrogram and pitch analyses (e.g., Figure 8(f)) indicate that during the breathy utterances in story readings, breathiness occurs more frequently in low-pitch intervals, while the voice quality tends to get back to modal phonation in high-pitch intervals. This rhythm pattern was observed in 90% of the "gentle/tender" tokens. This pattern is in contrast with the emphasized/excited speech in Figure 8(c), where breathiness is more prominent in high-pitch intervals than in low-pitch intervals.

Further, although this speaking style was predominant in story readings, it also appeared (with less frequency) when the speaker was talking to her child in a gentle/tender manner.

5.6. Confidential Talking, Talking-to-Oneself, Embarrassment, Diffidence. Confidential talking is often characterized by whisper or whispery voice over the whole utterance, being low-powered in comparison to the normal phonation (all "confidential talking" tokens). This speaking style also appeared when the speaker was thinking, embarrassed, or talking/asking to oneself (e.g., Figure 8(g)). In 55% of the tokens whispering occurred over the whole utterance, while in 25% of the tokens, it occurred at the end portions of the utterances.

Expressions of embarrassment and talking/asking to oneself include "*naniyattakkena*", "*nantsuttakke*", "*…wakannai*", "*dooshiyoo …*", which mean "I cannot remember…", "I'm not sure…", "what should we do?"

In utterances annotated as "diffidence", whispering often occurred at the end portion of the utterances (85% of the "diffidence" tokens).

5.7. Sighing Speech: Disappointment, Regret, Weariness, Relief. A couple of samples of sighing speech were found in the analysis data, expressing disappointment, regret, weariness and relief.

Sighing speech was mostly observed in the interjections "*aah*" and "*haa*", but was also observed along with speech utterances like in "*ah, arigatoo*" ("oh, thanks") (e.g., Figure 8(h)).

Sighing speech was often characterized by breathiness accompanied by a low decreasing pitch intonation (80% of the "sighing speech" tokens). However, an unvoiced whispered (or aspirated) quality was also observed in one of the tokens.

5.8. Laughing Speech: Funny Laughs, Bitter Laughs, Forced (Non-Spontaneous) Laughs. Laughing speech was often accompanied by a breathy (aspirated) voice quality. This was a common feature for almost all speakers.

Breathiness in laughing speech sounds different from the other items. One difference is that in laughing speech, the power of the voiced components also changes rhythmically, besides the breathy (aspirated) components, sounding like an alternation of the vowel sounds and the aspirated /h/.

Further, three types of laughs (funny laughs, bitter laughs, forced laughs) were identified. Although all types were characterized by breathiness (aspiration), preliminary observations indicated that in funny laughing speech, breathiness (aspiration) appeared rhythmically over the whole or part of the utterance (as in Figure 8(i)), while in bitter laughing speech, a strong breathiness (aspiration) tended to occur shortly, often at the end portion of the utterance (as in Figure 8(j)). The discrimination between bitter laughs and forced (non-spontaneous or social) laughs was more ambiguous, so that the context might be influencing. Detailed analysis for discrimination of different types of laugh is subject for future work.

6. Discussions

The acoustic analyses in Section 4 showed that the proposed acoustic parameters (*F1F3syn*, *breathiness power,* and *fricative power*) could potentially be used to characterize breathy/whispery segments. However, the *breathiness power* measure was found to have problems, mainly in the phoneme transition portions, where some peaks in the *breathiness power* contour were often observed (e.g., in Figures 8(a) and 8(e)), but are not particularly perceived as breathy. This is probably because the components in the F1 and F3 bands are unsynchronized in these transitional segments. Further constraints have to be considered to avoid misdetection of breathiness in such transitional segments. Also, although the *fricative power* (at 4000∼8000 Hz range) was effective to identify about 70% of fricative segments, misdetection occurred in some of the breathy segments. A better phonetic characterization, for example, by using MFCC parameters, could improve discrimination between breathy and fricative segments.

Although "breathy" and "whispery" types of phonations were not strictly separated in the present work, perhaps the introduced acoustic measures would be able to distinguish them, since the "true breathy" voices have softer aspiration noise compared with whispery voices. The relevance of a strict distinction of these two types of phonation in the expression of different paralinguistic information should also be investigated. These are subject for future work.

From the analysis results in Section 5, we can infer that breathiness occurring along with lower pitch and being closer to a whispered quality (as in real feeling expression, surprise/admiration, gentle/tender speaking style, confidential talking, diffidence) is more controlled and attitude-related, while the one occurring along with higher pitch (as in excitement, surprise/unexpectedness) is more spontaneous and emotion-related.

Regarding gender differences in the appearance of breathiness, analyses in the present data indicated that breathiness is much more common in female speakers (7.5% of the whole utterances) than in male speakers (2.0% of the whole utterances). This could be related to the physiological properties of male and female vocal folds, where males tend to have a more complete glottal closure compared with females [24]. Also, in female speakers, a large variety of paralinguistic information were found, while in male speakers, the breathy utterances were mostly frequent in laugh (about 53%), and in diffidence (about 16%).

Regarding the relationships between breathiness and the paralinguistic information conveyed by them, it is worth mentioning that breathiness is not strictly necessary for expressing a specific attitude or emotion, that is, the presence of breathiness may not serve as a "cue" for a specific attitude or emotion. Other strategies such as raising the pitch, or using other voice qualities (like pressed voices) could express the same attitudes or emotions expressed by breathiness. However, our analyses indicate that when breathiness appears, it is likely to express some attitudinal or emotional behavior of the speaker.

Regarding language dependency, we consider that the usage of intonation-related prosodic features and voice quality features may vary depending on the language, as stated in the introduction, so that part of the paralinguistic information items carried by breathy/whispery voices found in the present work might be specific for Japanese. However, similar methodologies could be applied for analyzing the appearance of breathiness in other languages.

Finally, although the present work focused on speech data of speakers with normal voice, similar analysis approaches could be applied also for studying pathological voices, by characterizing the temporal patterns of breathiness along the utterances.

7. Conclusion

The roles of breathy and whispery voices were analyzed in Japanese natural conversational speech of several speakers. Breathy and whispery voices were shown to appear with several dynamic patterns, expressing a variety of paralinguistic information.

Breathiness in low-pitch intervals, accompanied by a soft voice quality, appears in the expression of politeness, gentleness or tenderness, which can be considered as attitudinal behaviors of the speaker. Breathiness (whispery voice) in high-pitch intervals is more spontaneously produced, and often appears to express an excited emotional state of the speaker, such as happiness, surprise. Another type is when the whole or almost the whole utterance becomes whispered (unvoiced), appearing in confidential talking, embarrassment, or when the speaker is talking to oneself. A breathy voice quality also appears in sighing speech. In this case, the intonation has a lowering pattern with low pitch and a soft voice quality, expressing disappointment, regret, weariness, or relief. Finally, laughing speech is also characterized by breathiness (aspiration), and further acoustic analysis accounting other prosodic features would be necessary for their identification.

The acoustic parameters presented in the paper were shown to potentially characterize the breathy/whispery segments. However, improvements are still necessary, mainly in the phoneme transitions and in the discrimination with fricatives. Future works include improvement of the acoustic

features, identification of the different rhythmic patterns of breathiness, and mapping with paralinguistic information items.

Acknowledgment

This work was partly supported by the Ministry of Internal Affairs and Communications and by the Ministry of Education, Culture, Sports, Science and Technology.

References

[1] J. Laver, "Phonatory settings," in *The Phonetic Description of Voice Quality*, pp. 93–135, Cambridge University Press, Cambridge, UK, 1980.

[2] D. Erickson, "Expressive speech: production, perception and application to speech synthesis," *Acoustical Science and Technology*, vol. 26, no. 4, pp. 317–325, 2005.

[3] C. T. Ishi, H. Ishiguro, and N. Hagita, "Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality," *Speech Communication*, vol. 50, no. 6, pp. 531–543, 2008.

[4] M. Gordon and P. Ladefoged, "Phonation types: a cross-linguistic overview," *Journal of Phonetics*, vol. 29, no. 4, pp. 383–406, 2001.

[5] G. Klasmeyer and W. F. Sendlmeier, "Voice and emotional states," in *Voice Quality Measurement*, pp. 339–358, Singular Thomson Learning, San Diego, Calif, USA, 2000.

[6] C. Gobl and A. Ní Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Communication*, vol. 40, no. 1-2, pp. 189–212, 2003.

[7] H. Kasuya, M. Yoshizawa, and K. Maekawa, "Roles of voice source dynamics as a conveyer of paralinguistic features," in *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP '00)*, pp. 345–348, 2000.

[8] M. Fujimoto and K. Maekawa, "Variation of phonation types due to paralinguistic information: an analysis of high-speech video images," in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS '03)*, pp. 2401–2404, 2003.

[9] M. Ito, "Politeness and voice quality—the alternative method to measure aspiration noise," in *Proceedings of the 2nd International Conference on Speech Prosody*, pp. 213–216, 2004.

[10] J. Kreiman and B. Gerratt, "Measuring vocal quality," in *Voice Quality Measurement*, pp. 73–102, Singular Thomson Learning, San Diego, Calif, USA, 2000.

[11] K. Stevens, "Turbulence noise at the glottis during breathy and modal voicing," in *Acoustic Phonetics*, pp. 445–450, The MIT Press, Cambridge, Mass, USA, 2000.

[12] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[13] H. M. Hanson, "Glottal characteristics of female speakers: acoustic correlates," *Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 466–481, 1997.

[14] P. Alku and E. Vilkman, "Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering," *Speech Communication*, vol. 18, no. 2, pp. 131–138, 1996.

[15] D. Michaelis, T. Gramss, and H. W. Strube, "Glottal-to-noise excitation ratio—a new measure for describing pathological voices," *Acustica*, vol. 83, no. 4, pp. 700–706, 1997.

[16] T. Ohtsuka and H. Kasuya, "Aperiodicity control in ARX-based speech analysis-synthesis method," in *Proceedings of the 7th European Conference on Speech Communication and Technology (EUROSPEECH '01)*, pp. 2267–2270, September 2001.

[17] C. T. Ishi, "A new acoustic measure for aspiration noise detection," in *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP '04)*, vol. 2, pp. 941–944, 2004.

[18] M. R. Schroeder, "Hilbert envelope and instantaneous frequency," in *Computer Speech: Recognition, Compression, Synthesis*, pp. 174–177, Springer, New York, NY, USA, 1999.

[19] N. Campbell, "Databases of Emotional Speech," in *Proceedings of ISCA (International Speech Communication and Association) ITRW on Speech and Emotion*, pp. 34–38, 2000.

[20] The Corpus of Spontaneous Japanese, http://www.kokken.go.jp/katsudo/seika/corpus/public.

[21] K. R. Scherer and H. Ellgring, "Multimodal expression of emotion: affect programs or componential appraisal patterns?" *Emotion*, vol. 7, no. 1, pp. 158–171, 2007.

[22] S. Toki and M. Murata, "Pronunciation & task learning—Japanese for foreigners," *Atake Shuppan*, pp. 19–35, 1987 (Japanese).

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

[24] L. A. Rammage, R. C. Peppard, and D. M. Bless, "Aerodynamic, laryngoscopic, and perceptual-acoustic characteristics in dysphonic females with posterior glottal chinks: a retrospective study," *Journal of Voice*, vol. 6, no. 1, pp. 64–78, 1992.