ISCA Archive
http://www.isca-speech.org/archive

Speech Prosody 2008
Campinas, Brazil
May 6-9, 2008

The roles of breathy/whispery voice qualities in dialogue speech

Carlos Toshinori Ishi, Hiroshi Ishiguro & Norihiro Hagita

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

Abstract

Qualitative analyses are conducted in spontaneous dialogue speech of several speakers, to verify the paralinguistic roles of breathy/whispery voice qualities in communication. Analyses show that breathy/whispery voices carry a variety of emotion- or attitude-related paralinguistic information. Breathiness often appeared in emphasized words/phrases, having the effect of calling/catching the listener's attention. It also rhythmically appeared in utterances expressing the speaker's excitement. In backchannels, breathiness has the effect of expressing politeness or interest to the listener's talk. When accompanied by a softer voice quality, breathiness is used to call/catch the listener's attention, while expressing gentleness or tenderness. A more whispered and low-powered voice quality appears in confidential talking, embarrassment, or when the speaker is talking to oneself.

1. Introduction

Besides the linguistic information, the understanding of paralinguistic information is also important in spoken dialog systems. Although prosodic features, 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 (caused by non-modal phonations, such as breathy, whispery, creaky and harsh [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 prosodic 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.

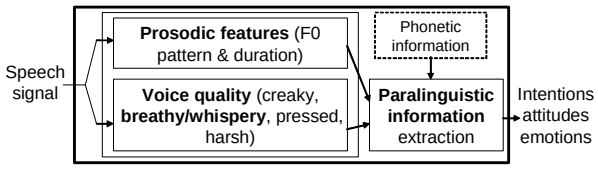


Figure 1: *A framework for paralinguistic information extraction considering prosodic and voice quality features.*

In the present work, we focus on the breathy/whispery voice qualities, and analyze their communication roles (i.e., the variations in paralinguistic information) in spontaneous dialogue speech, for several speakers.

Regarding the terminologies, "breathy" and "whispery" phonations can distinctly be defined from a physiological viewpoint [1]. 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. In breathy phonation, the constriction results from an incomplete closure of the vocal folds during its vibration. Both phonations are characterized by an air escape through the glottis, which increases the noisy (non-harmonic) components in the frequency bands around the third formant [4,5].

Breathy and whispery voices have been reported 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 [6]. In [7,8], relationships between different phonation types and paralinguistic information like emotions and attitudes are reported. In [7], whispery voice was found in "fear" while breathy voice was found in "sad" voice. Correlations between synthesized breathy/whispery voices and perception of relaxed/stressed, sad/happy, and intimate/formal are also reported in [8]. Breathiness is also reported to appear in the expression of disappointment [9,10]. In Japanese spontaneous speech, possible use of breathiness for expressing manner or politeness is reported [11].

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 [12]. 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 an auditory impression of turbulent noise (aspiration noise) caused by an air escape through the glottis.

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

Before finishing the introduction, it is worth mentioning about a number of parameters which try to characterize the acoustic features 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] characterize the spectral slope properties of breathy voice, while *GNR* (glottal-to-noise ratio) [15], $f_{aperiodic}$ (boundary frequency between harmonic and aperiodic components) [16] and *F1F3syn* (synchronization of the amplitude envelopes of the first and third formant frequency bands) [17] try to reflect the effects of the noise components caused by air escape through the glottis. In the present work, qualitative analysis is conducted based on auditory impression and spectrogram analysis, so that a quantitative analysis based on acoustic parameters is left for future work.

2. Description of the speech data for analysis

Two databases of spontaneous speech were analyzed in the present work. One is the JST/CREST ESP expressive speech database [18]. Data of eight speakers (six female and two male speakers) were selected for analysis. This database can be classified in two types:

- FAN, FYM, FSM, FYS (female, 30s): natural daily conversations (including telephone calls) between family members, friends, and non-familiar people (hospital, companies). The length of each dialogue file varies from 10 to 30 minutes.
- JFA (female, 40s), JFB (female, 30s), JMA (male, 20s), JMB (male, 30s): 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 [19]. This corpus is constituted by monologue and dialogue speech data. For the present work, dialogue data was analyzed. The dialogues are between speakers which are familiar and not familiar with each other. Each dialogue has approximately 10 minutes.

- D01 (16 dialogues; total 3.2 hours): interview after simulated public speaking.
- D02 (16 dialogues; total 3.1 hours): task-oriented dialogues.
- D03 (16 dialogues; total 3.6 hours): free dialogue conversations.
- D04 (10 dialogues; total 2.1 hours): interview after academic presentations.

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

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

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, including breathy, whispery, non-voiced whisper, and aspirated sounds at utterance finals. One of the subjects identified Br/Wh voices in 1584 utterances, while the other subject identified in 1752 utterances. The 1134 utterances, where both subjects identified the presence of Br/Wh voices, were used for subsequent analysis.

3. Identified paralinguistic information (PI) carried by breathy/whispery voices

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], [9-11]) was given to the subjects, but items were allowed to be freely added, according to the subject's impression. The first set of PI items included: 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 the following items: forced laugh, bitter laugh, excitation, emphasis, calling for attention, interest, gentleness, real feeling expression, sympathy, keenness, emotion quoting, diffidence, undecided, and talking-to-oneself.

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

The distribution of the PI items where matching was obtained after revision by the three subjects are shown in Fig. 2. The PI items could be summarized as follows.

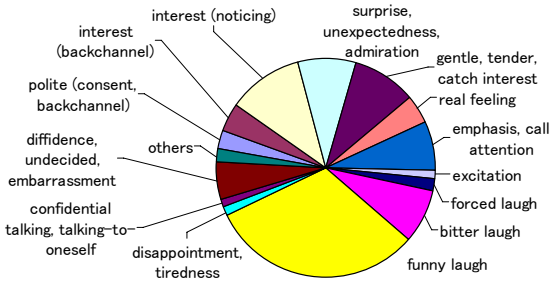


Figure 2: *Distribution of the paralinguistic information carried by breathy/whispery voices.*

A) Expression of emphasis, attention

There are a number of ways for emphasizing a word/phrase while speaking. One is raising the pitch in the focused word/phrase, another is increasing the power, and another is lengthening the word [20].

The analyses of the present spontaneous speech database have shown that breathiness often appeared in emphasized word/phrases, and has the effect of calling/catching the attention of the listener. In some of the utterances the only use of breathiness, without a significant raise in pitch or power, was effective for expression of emphasis. Breathiness was frequent in the phrase beginnings, high-pitch accent portions, but also appeared along the whole phrase.

A.1) Expression of real feelings

Some of the utterances annotated as "emphasis" were also interpreted as if the speaker was expressing a "real feeling". For example, the utterance "kore wa futoi yo" ("this is thick") accompanied by Br/Wh voice 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, as reported in [21]. The effects of real feeling expressions seem to be stronger as the speaking style approximates to whispered speech (absence of voicing).

A.2) Excitement

Along with "emphasis", some of the utterances were annotated as "excited". Br/Wh segments rhythmically appeared within and across the utterances, when the speaker was speaking in an excited state. Fig. 3a shows an excited speech utterance containing Br/Wh segments. Note that the harmonic components are present in the low frequency

components around 0 to 1500 Hz (horizontal lines), but they become weaker or absent in the range around 1500 to 4000 Hz, in the Br/Wh segments.

Changes in voice quality often 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.

B) Expression of politeness/interest.

Politeness and interest were annotated in most of the interjections "hai" and "un", where Br/Wh was perceived. Both interjections are commonly used as backchannels in dialogue when spoken by a falling intonation. "hai" ("yes") is usually a more formal backchannel, while "un" ("uh han") is a casual backchannel.

Results showed that breathy "hai" utterances are perceived as more formal/polite than the non-breathy "hai" utterances, while breathy "un" utterances express more interest in the interlocutor's talk, than their non-breathy counterparts.

The short interjection "ah", which express noticing, was commonly followed by backchannels "soo"/"soodesuka", and also often accompanied by breathiness. Breathy "ah" is thought to express more interest than the non-breathy ones.

C) Expression of surprise/unexpectedness/admiration.

Two types of speaking styles were frequent in the expression of surprise, unexpectedness and admiration. One is an excited mode, as in the excited speech of item A.2), which is though to be more emotion-related, while the other is a quiet mode, expressing interest in the interlocutor's talk, and is thought to be a more attitude-related behavior of the speaker.

When the excitement level of the speaker increases, a harsh voice quality (due to aperiodic vibrations of the vocal folds) tends to appear along with breathiness.

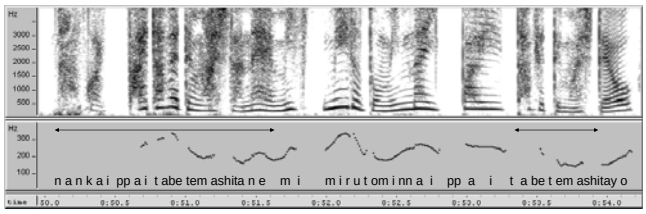
Almost all utterances annotated as surprise, admiration or unexpectedness, are accompanied by interjections or interjectional expressions like "eeh!", "sugoi!", "hontoo!", "hee", "haa", "waa", "ah!", "soonandesuka!", "soonanda!", "naruhodo", "uso!", which can equivalently be translated as "wow!", "really!?", "amazing!", "you're kidding!". Such linguistic information is thought to be important to discriminate them from the excited speech of A.2.

D) Expressivity in story readings: calling attention, gentleness/tenderness.

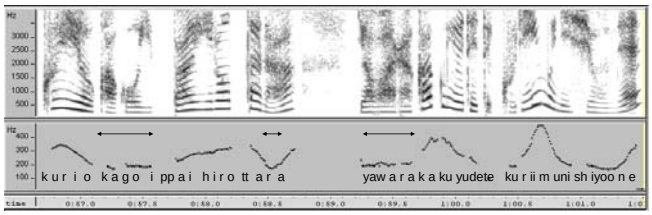
Breathiness accompanied by a soft voice quality appeared rhythmically along the utterances of the speaker FYM (mother), when she was reading stories to her child (FYM001_04, FYM001_07, FYM001_27). This change in voice quality (from the speaker's normal speaking voice) displays expressivity and is thought to have the effect of calling/catching the attention of the listener, while expressing gentleness/tenderness.

Regarding the rhythm of breathiness within an utterance, spectrogram analysis (e.g. Fig. 3b) indicates 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 is in contrast with the emphasized/excited speech in Fig. 3a, where breathiness occurs across the whole word or phrase, including high-pitch intervals.

From the present results, one can say that breathiness occurring along with low pitch (as in the gentle/tender speaking style) is more controlled and attitude-related, while



(a) Emphasis/excitement.



(b) Gentle/tender (story reading)

Figure 3: *Examples of spectrograms and pitch contours of spontaneous speech utterances including breathy/whispery voice (indicated by arrows).*

the one occurring along with higher pitch (as in the excited speaking style) is more spontaneous and emotion-related.

E) Confidential talking, talking-to-oneself, embarrassment, diffidence.

Confidential talking is often whispered or whispery-voiced over the whole utterance, being low-powered in comparison to the normal phonation. A similar speaking style appears when the speaker is thinking, embarrassed or talking/asking to oneself. In the utterances annotated as "diffidence", the whispering occurred often at the end portion of the utterances.

Expressions of embarrassment include "naniyattakkena", "nantsuttakke", "… wakannai", which mean "I can't remember …".

F) Sighing while speaking: disappointment, regret, weariness, relief.

A couple of samples of sighing speech were found in the analysis data. Sighing speech was often characterized by breathiness accompanied by a low decreasing pitch intonation, signaling the speaker's disappointment, regret, weariness or relief. The interjections "aah" and "haa" were representative of sighing speech.

G) Laughing while speaking: 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 all 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, while in bitter laughing speech, a strong breathiness (aspiration) tended to occur only in the very end portion of the utterance. However, the discrimination between bitter and forced laughs was more ambiguous, so that

the context might be influencing. More detailed analysis is necessary for identification of different types of laugh.

4. Discussions

Regarding gender differences, analyses in the present work indicated that Br/Wh voices are much more common in female speakers (7.5% of the utterances) than in male speakers (2.0% of the utterances). Among the Br/Wh utterances in male speakers, about 53% appeared in laugh, and about 16 %, in diffidence.

Regarding the relationships between Br/Wh voices and the paralinguistic information conveyed by them, it is worth mentioning that such voice qualities are not strictly necessary for expressing a specific attitude or emotion, however, when Br/Wh voices appear, they are likely to express some attitudinal or emotional behavior of the speaker. Other voice qualities, such as pressed voice, could be used instead for expressing the same attitude expressed by Br/Wh voices (e.g., in real feeling expression).

Finally, regarding language dependency, we consider that the usage of prosodic 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 are specifically for Japanese. Further analyses on spontaneous dialogue would be necessary to verify the usage of Br/Wh voices in other languages.

5. Conclusions

The roles of breathy/whispery voices during natural conversational speech of several speakers were analyzed. Breathy/whispery voices were shown to appear in 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 is necessary for their identification.

The next step of the present work is a quantitative analysis of acoustic features for characterization and identification of the different rhythmic patterns of breathiness, and their mapping with paralinguistic information items.

6. Acknowledgements

This research is partly supported by the Ministry of Internal Affairs and Communications and the Ministry of Economy Trade and Industry. We thank Kyoko Nakanishi and Maiko Hirano for the valuable helps in the data annotation and interpretation.

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] Ishi, C.T., Ishiguro, H. and Hagita, N., 2006. Using prosodic and voice quality features for paralinguistic information extraction. CD-ROM Proc. of *The 3rd International Conference on Speech Prosody*.

[4] Stevens, K., 2000. Turbulence noise at the glottis during breathy and modal voicing. In *Acoustic Phonetics*, The MIT Press, 445-450.

[5] Klatt, D., Klatt, L., 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoustic. Soc. Amer.*, Vol. 87: 820-857.

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

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

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

[9] Kasuya, H., Yoshizawa, M., Maekawa, K., 2000. Roles of voice source dynamics as a conveyer of paralinguistic features. Proc. *ICSLP 2000*, 345-348.

[10] Fujimoto, M., Maekawa, K., 2003. Variation of phonation types due to paralinguistic information: An analysis of high-speech video images. Proc. *15th ICPhS*, 2401-2404.

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

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

[13] Hanson, H., 1997. Glottal characteristics of female speakers: Acoustic correlates. *J.Acoustic. Soc. Amer., Vol. 101*: 466-481.

[14] Alku, P., Vilkman, E., 1996. Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering. *Speech Communication*, Vol. 18, No. 2, 131-138.

[15] Michaelis, D., Gramss, T., Strube, H.W., 1997. Glottal-to-noise excitation ratio – a new measure for describing pathological voices. *Acustica*, Vol.83, 700-706.

[16] Ohtsuka, T., Kasuya, H., 2001. Aperiodicity control in ARX-based speech analysis-synthesis method. Proc. *Eurospeech 2001*, 2267-2270.

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

[18] JST/CREST Expressive Speech Processing homepage, http://feast.atr.jp/esp/esp-web/index.html

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

[20] Toki, S., Murata, M., 1987. *Pronunciation & task learning – Japanese for foreigners*. Atake Shuppan, 19-35. (In Japanese)

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