Paralinguistic Effects on Voice Quality: A Study in Japanese

Caroline Menezes & Kikuo Maekawa

National Institute for Japanese Language, Japan
{menezes;kikuo}@kokken.go.jp

Abstract

This study analyzes two spectral properties in vowel segments, H1-H2 (related to glottal opening) and H1-A3 (related to the speed of vocal fold closing gesture) in an attempt to infer the voice quality variation associated with different types of paralinguistic information (PI) types. Results suggest that both glottal opening and closing speed of the glottis differ significantly depending on PI. However, for some PI types there were also significant syllable effects. The correlation between F0 and these two voice parameters was very low leading to the conclusion that just F0 differences cannot account for the observed voice quality variation. Significant differences were also noted for the power of speech waveform (RMS) according to PI. Inter-speaker variation was noted especially for 'suspicion'.

1. Introduction

In recent times speech scientists have focused their attention on the effects of emotions and attitudes with respect to various acoustic parameters of voice signals. Emotional attributes of speech signals and paralinguistic information in general both differ from linguistic information in that they are expressed, mostly, by prosodic features rather than segmental features. According to Fujisaki [1], paralinguistic information is conveyed by the speaker to the listener intentionally, while emotion is conveyed in an involuntary process.

Previous work by the second author revealed systematic differences in phonetic features like duration, pitch contour, and voice quality for varying PIs [2, 3, and 4]. Also articulatory analyses (on the current data set) show that there is a tendency for the forward displacement of the tongue dorsum for suspicious utterance when compared with that of admiration, which shows a backward displacement compared with neutral. These results correlate strongly with raising and lowering of F2 (the second formant frequency) values [2, 3]. However, the physiological displacement of the tongue was seen in the entire utterance including both consonants and vowels, leading to the speculation that this displacement is not just the manipulation of segmental features, but rather of the entire utterance [3]. Moreover, high-speed digital video imaging studies showed that glottal area and glottal waveform were characteristically different for 'disappointment', 'suspicion' and 'neutral' for all segments of a single word utterance [5]. These results call for more detailed analysis of voice quality differences in paralinguistic information.

In this paper we conducted a systematic analysis on the acoustic recording, in order to infer what voice quality parameter is salient in distinguishing different PI types namely, 'neutral', 'admiration', 'suspicion' and 'disappointment'.

2. Data

2.1. Recording

The data used in this study are part of a larger data set collected for the purpose of studying the articulatory gestures involved in paralinguistic information. This data were recorded using the EMA (Carstens 100) system (NTT basic research laboratory, Atsugi, Japan). In this study, however, only the acoustic signals were analyzed. The data analyzed consisted of two speaker's production of the short phrase /sasadaga/ (surname 'Sasada' followed by nominal case particle 'ga') with four different PI types namely, 'neutral', 'admiration', 'suspicion' and 'disappointment'. Both speakers were native male speakers of standard Japanese. The instructions given to the subjects about the four PI types can be found in the CD-ROM version of [4]. This phrase, uttered in isolation, was selected as it contained the same vowel in all syllable positions in the phrase, thus avoiding vowel influences on voice quality. In addition the phrase has no lexical pitch accent on any of the syllables, and pitch changes that occurred from syllable to syllable relatively manifested the direct effect of PI. For one speaker there were twelve repetitions per each PI and for the other there were 16 repetitions. The acoustic signal was digitized at the sampling frequency of 16000Hz with a16-bit quantization.

2.2. Analysis

Spectral analyses were made using the Wavesurfer program (developed at the Royal Institute of Technology, Sweden). For every syllable in the phrase /sasadaga/, a point near the center of the acoustic vowel segment was selected for the spectral measurements. This point, in our data, coincided with relatively stable formants. The following spectral measurements were made:

- H1-H2 (difference in dB amplitude between the first and second harmonics)
- H1-A3 (difference in dB between the first harmonic amplitude H1 and the amplitude of the peak harmonic in the F3 region).

Fundamental frequency (F0) and RMS were also estimated at the same point in time as the spectral measurements.

3. Results

3.1. H1-H2 Measurement (Glottal Opening)

The extent of glottal opening during the glottal cycle affects primarily the lowest frequency components of speech signals and can be approximately represented by the difference between the amplitudes of the first two harmonics (H1-H2) in

the spectrum [6 & 8]. For example, according to Stevens & Hanson, a change from 30% to 70% in open quotient results in a 10dB difference in the H1-H2 value [6].

Figure 1 shows the mean H1-H2 values for the different PI types plotted separately for each speaker.

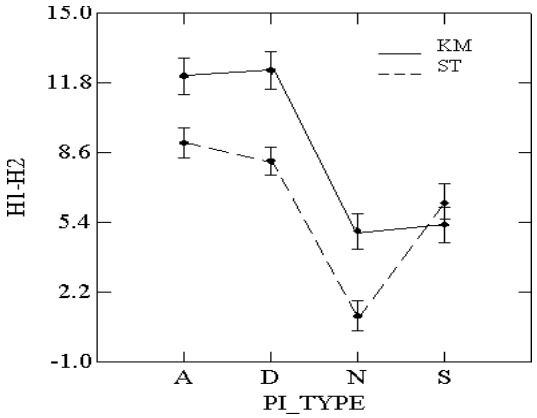

Figure 1: *Mean H1-H2 (dB) differences separated by speaker. A, D, N, S = admiration, disappointment, neutral and suspicion respectively. KM, ST are the two speakers. Error bars represent ± 1SD.*

Speakers were not pooled as there were differences in ranges. Speaker ST had a smaller range (11dB) when compared to speaker KM (13dB). The lower mean values for speaker ST indicate that this speaker had a comparatively closed glottis than speaker KM.

Both speakers had a significantly smaller glottal opening for 'neutral' utterances when compared to 'admiration' and 'disappointment' ($p<0.001$). There was no significant difference between 'admiration' and 'disappointment'. 'Suspicion' was produced with a smaller glottal opening than 'admiration' and 'disappointment', but larger than 'neutral'. However, this distinction in glottal opening between 'neutral' and 'suspicion' was only significant for speaker ST ($p<0.001$). This was one instance of speaker variation.

3.2. H1-A3 Measurement (Spectral Tilt)

Voice quality is not only affected by glottal opening (H1-H2) but also by the "abruptness" of glottal approximation in the closing phase, *i. e*., the speed of vocal fold movement in the glottal closing phase [6]. This effect is manifested in the middle and high frequency components, thus affecting the tilt in the spectrum.

To analyze the rate of glottal closure, we measured the difference between the first harmonic and the strongest harmonic at the F3 region (H1-A3). Roughly, this is a measure of the tilt of the spectral envelope from the fundamental component toward higher frequencies. Figure 2 plots the mean H1-A3 values for the different PI types separated by speakers. We see that speaker KM had larger tilt values when compared to ST, indicating a more gradual anterior to posterior closing movement of vocal folds. In general, both speakers had significantly larger tilt for 'admiration' and 'disappointment' when compared with 'neutral' and 'suspicion' ($p<0.001$). KM on average had a less sharp tilt for 'suspicion' when compared to all other PI types including 'neutral' (significant for A *vs*. D at $p<0.001$). This would indicate an increase in the high frequency amplitudes reflecting a rather abrupt decrease of flow in the closing phase. On the other hand, for the same PI type, ST showed low spectral tilt but it was steeper than the 'neutral' condition.

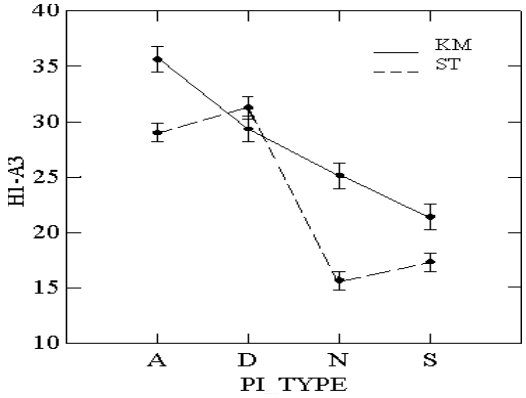

Figure 2: *Mean H1-A3 (dB) differences separated by speaker. A, D, N, S = admiration, disappointment, neutral and suspicion respectively. KM, ST are the two speakers. Error bars represent ± 1SD.*

3.3. Syllable Position Effects

Even though one-way ANOVA showed significant difference for the different PI types for both glottal opening and spectral tilt, there were also significant differences across syllables (despite the use of the same vowel /a/) for all PI types.

Table 1 lists the average (± 1SD) values of H1-H2 and H1-A3 for all syllables within each PI type separated by speaker. Syllable position in the phrase is indicated as '1', '2', '3' and '4' and PI as 'N', 'A', 'S' and 'D'. The increasing number of * indicate which and to what degree each syllable position was significant ($p<0.05$). Single * indicates significant difference with one other syllable in the same PI, and *** indicates significant differences with all other syllables in the same PI. Two ** indicates significance with two other syllables in the same PI.

In this table we see that syllable effect is prevailing except for 'disappointment' for speaker KM. This suggests that syllable effects are very strong even in this short phrase 'Sasadaga'. The proportionately large number of *** cells (10:16) for 'syllable 1' show that this utterance-initial syllable is characteristically different from all succeeding syllables. A couple of explanation can be postulated: first, the first syllable (unless it is accented) in a Japanese phrase is uttered in low tone, and in our example phrase, there is no pitch accent phonologically specified (and therefore there is no other low tone associated with any of the syllables in the phrase). Second, the vowel in the first syllable is surrounded by the voiceless consonant /s/ and therefore we assume that approximation of the glottis tends to be incomplete under the effect of co-articulation. Speech intensity values for this syllable are very low seeming to confirm this supposition, see Table 2.

In the case of 'suspicion' we see there are generally large differences between most syllable positions for both speakers (8:16 *** cells). Speakers had a tendency to produce the first syllable in creaky or harsh voice but the final syllable in falsetto. The reason for this is the rather sharp and steady

increase in fundamental frequency culminating with very high values on the final syllable. The intonation pattern for the PI types analyzed in this paper has been previously reported (see [2], [3], and [4] for details).

3.4. Influence of Fundamental Frequency

In order to evaluate the mutual dependence between the voice fundamental frequency and spectral amplitude characteristics that are assumed to reflect the vocal fold vibration pattern that we measured, we conducted a correlation analysis between F0 and H1H2, and F0 and H1H3. The influence of F0 on the two voice quality parameters was very low: R^2 = 0.05, 0.15 for H1H2 and H1A3, respectively ($p<0.001$). The data were pooled for both speakers. In Table 2 we give the range, the

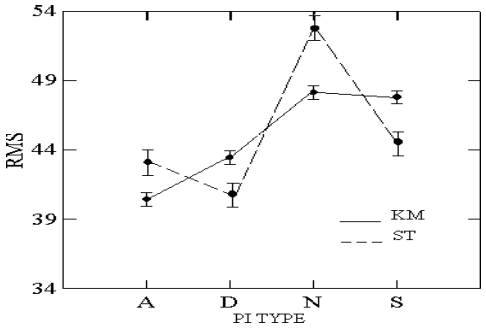

Figure 3: Mean RMS separated by speakers KM, ST. A, D, N, S = admiration, disappointment, neutral and suspicion respectively. Error bars represent ± 1SD.

Table 1: *Mean H1-H2, H1-A3 values (SD) separated by speaker, syllable position (1, 2, 3, 4) and PI types. N=neutral, A=admiration, S=suspicion and D=disappointment. Shading indicates significance at p<0.05.*

Speaker	PI	H1-H2				H1-A3			
		1	2	3	4	1	2	3	4
KM	N	9.94 (6.5)***	4.13 (5.9)*	3.89 (5.3)*	2.79 (5.4)*	28.24 (6.7)***	24.39 (6.4)**	21.60 (5.7)***	23.90 (6.8)**
	A	9.78 (4.8)*	18.28 (8.6)**	12.43 (8.4)*	8.02 (2.3)*	23.36 (9.8)***	40.96 (7.9)*	38.47 (9.3)*	38.80 (4.1)*
	S	-3.32 (2.8)***	8.61 (4.6)*	7.90 (3.8)*	8.68 (1.8)*	2.93 (4.2)***	20.41 (6.1)***	26.57 (5.4)***	32.83 (2.7)***
	D	13.39 (5.7)	13.57 (7.0)	10.87 (3.6)	11.81 (6.2)	30.96 (6.0)	29.60 (7.1)	27.24 (6.4)	27.82 (11.7)
ST	N	5.38 (1.8)*	8.71 (3.9)*	2.81 (4.9)**	10.53 (3.4)**	25.60 (2.8)*	21.78 (5.0)**	13.86 (4.0)***	29.13 (5.1)**
	A	1.71 (1.6)**	7.08 (3.4)**	1.53 (1.7)**	8.90 (6.3)**	14.92 (2.4)***	36.46 (3.5)***	18.91 (3.2)***	31.52 (4.0)***
	S	-0.23 (2.0)***	12.03 (3.1)***	6.01 (3.6)**	4.99 (3.5)**	14.94 (4.3)***	31.05 (3.9)**	22.36 (4.6)***	33.70 (4.2)**
	D	-2.58 (1.1)***	6.55 (2.9)**	13.48 (3.1)***	6.67 (4.2)**	13.28 (2.9)***	28.70 (3.4)**	19.97 (3.5)***	31.38 (4.1)**

Table 2: *F0 values pooled for speakers and syllable position and separated by PI types.*

	Neutral	Admiration	Suspicion	Disappointment
N	112	112	112	112
Min	86.000	73.000	59.000	90.000
Max	164.000	230.000	344.000	162.000
Mean	120.964	164.013	153.054	109.464
Std. Dev	17.393	46.589	74.128	10.804

mean and standard deviation of F0 pooled across speakers and syllable positions but separated by PI types. From this table we can see that F0 range was large for 'suspicion' and 'admiration' but small for 'neutral' and 'disappointment'.

3.5. RMS of Speech Waveform

While analyzing the acoustic signal, noticeable differences among PI types were observed in the amplitude of waveform. As one simple measure of waveform characteristics, RMS value was computed using a Hamming window with a 0.01 sec. time frame. Figure 3 plots the mean RMS for different PIs separated by speaker.

One-way ANOVA showed that both speakers had significantly different speech intensity depending on PI types ($p<0.001$). Further, post-hoc tests revealed that 'neutral' utterances were produced with significantly more mean intensity than 'admiration' and 'disappointment' for both speakers. There was also significant difference in intensity between 'neutral' and 'suspicion' for ST. In this respect, it is interesting to note that this speaker also had a larger mean glottal opening (Figure 1) when compared to KM.

Syllable effects were also observed for both speakers. Table 3 lists the mean RMS values separately for the two speakers and syllable position. Here the data is pooled for all PI types. We see that mean RMS values vary depending on syllable position, the initial syllable having the lowest RMS value compared to other syllable positions.

Table 3: *Mean RMS (±1 SD) separated by speaker and syllable position and pooled for PI types.*

Speaker	1	2	3	4
KM	39.49 (6.08)	44.49 (7.84)	49.39 (5.36)	48.15 (5.41)
ST	42.55 (4.78)	45.37 (6.00)	46.98 (4.39)	44.89 (3.57)

4. Discussion

These results showed a clear differentiation in voice source characteristics as a function of speaker attitudes, which is considered part of paralinguistic information. Of the four types studied here, *i.e.*, 'neutral', 'admiration', 'suspicion' and 'disappointment', the glottis was relatively more open for 'admiration' and 'disappointment' than for 'neutral' and 'suspicion' in the two native Japanese subjects we examined. The concomitant large H1-H2 and H1-A3 values would indicate the presence of a glottal chink for 'admiration' and 'disappointment'. A digital video imaging study pertaining to the same subject KM has also shown that the vocal processes remained apart during the closed phase throughout an entire word for 'disappointment' utterances, and this phonation was considered breathy [5]. 'Admiration' was not reported in this study.

In 'suspicion' we saw that the both male speakers varied considerably. The digital video imaging study mentioned above also claimed that the phonation used in 'suspicion' was comparable to the creaky phonation, where there is approximation of the false vocal folds in addition to the true vocal folds [5]. This kind of phonation could be used to explain the pattern of speaker KM, who produces 'suspicion' with a less sharp spectral tilt and a small open quotient. Speaker ST produced 'suspicion' with a less sharp tilt, though not significantly different from 'neutral', but the open quotient was significantly larger. This significantly larger open quotient resulted also in low mean speech power (RMS). It could therefore be that this speaker used a phonation type different from the speaker KM for expressing suspicion.

Significant syllable effects were observed for both speakers for most of the investigated PIs. The initial syllable seems to be weak in speech signal intensity, and this may be in part due to the phonetic context of the vowel, which is surrounded by voiceless frication gestures or due to the less adducted glottis (larger open quotient). Further, due to the phrasal property of Tokyo Japanese, the initial syllable is pronounced in a low tone (the first syllable not being accented in our material). This syllable sounded harsh, and some breath noise may be present. It also had irregular F0.

While pitch change is a salient aspect of controlled voice quality (see Fujimura [9]) we saw no strong trend for F0 and either of the voice quality parameters for the different PI types. This absence of correlation was found for both speakers we studied here. Earlier studies have reported specific intonation contours for each PI type [2, 3, & 4]. Further, it was shown that listeners used pitch change (intonation) to judge PI type [2]. While conducting the spectral analyses, it was noted that in some cases the rising intonation in 'suspicion' resulted in a falsetto phonation in the final syllable. Also in some cases creaky phonation was observed in the initial syllable associated with a very low tone.

Lastly, only some of the syllable effects have been explained here; a more detailed study on syllable effects has to be conducted.

5. Conclusions

We can conclude that, PI types are characterized by voice quality as well as other phonetic attributes that were reported earlier [2 & 5]. Further research is necessary to investigate the effects of syllable position. Do listeners perceive PI type by the presence or absence of a particular voice quality over the entire utterance? Do listeners attend to a voice quality contour, paying attention to local changes within an utterance, when they distinguish paralinguistic information? These are questions that remain for future studies.

From this study it appears that one simple measure of voice quality might not be sufficient to distinguish all PI types. Of the two speakers used in this study each used different vocal strategies particularly for 'suspicion'. Perceptual effects of the two voice quality parameters need to be studied to understand which might be more salient to the listener.

6. Acknowledgement

We would like to thank Osamu Fujimura and Donna Erickson who have helped through many useful discussions and reviews of this paper. This study was supported in part by a Grant-in-Aid for Scientific Research for JSPS postdoctoral fellows (2005-7): 17-05246) to the first author.

7. References

[1] Fujisaki, H., 1997. Prosody, models, and spontaneous Speech. In *Computing Prosody*, Sagisaka et al. (ed.), Springer, 27-40.

[2] Maekawa, K., 2004. Production and perception of 'Paralinguistic' information. In *Proceedings of Speech Prosody*, Nara, Japan, 367-374.

[3] Maekawa, K.,; Kagomiya, T., 2000. Influence of paralinguistic information on segmental articulation. In *6th International Conference on Spoken Language Processing*, Beijing, China, v2, 349-52.

[4] Maekawa, K., 1998. Phonetic and phonological characteristics of paralinguistic information in spoken Japanese. In *5th International Conference on Spoken Language Processing*, Sydney, Australia, v2, 635-38.

[5] Fujimoto, M.; Maekawa, K., 2003. Variation of phonation types due to paralinguistic information: An analysis of high-speed video images. In *15th International Congress of Phonetic Sciences*, Barcelona, Spain, 2401-04.

[6] Stevens, K.N; Hanson, H.M., 1998. Classification of glottal vibration from acoustic measurements. In *Vocal Fold Physiology*, O. Fujimura and M. Hirano (ed.s). Cambridge MA: Hilltop University Press, 147-70.

[7] Buder, E., 2000. Acoustic analysis of voice quality: A tabulation of algorithms 1902-1990. In *Voice quality measurement*, R. D., Kent; M. J. Ball. Singular: San Diego, CA, 119-245.

[8] Laver, J., 1994. *Principles of phonetics*. Cambridge University Press.

[9] Fujimura, O. 2004. Prosody: A generalized concept. In G. Fant, H. Fujisaki, J. Cao & Y. Xu (*eds.*), *From Traditional Phonology to Modern Speech Processing*. Beijing: Foreign Language Teaching and Research Press., pp. 97-110.