Politeness and Voice Quality
– The Alternative Method to Measure Aspiration Noise

Mika Ito

Department of Theoretical and Applied Linguistics
University of Edinburgh, United Kingdom
mika@ling.ed.ac.uk

Abstract

This paper discusses some problems regarding the measurement of breathiness directly from the acoustic waveform, especially the estimation of aspiration noise found in the high frequency region, which is a prominent feature of breathiness.

Klatt and Klatt (1990) suggested the noise rating method for this, which employed the subjective ratings of visual observation of the irregularity of waveforms after they were band-pass filtered around the third formant frequency (F3), so as to quantify the aspiration noise of higher frequency region. However this method heavily depends on individual raters' subjective observation of the waveform. It is therefore questionable if the ratings are reliable.

Since the interest of this study is to examine the correlation between breathiness and politeness, the technical problem of this noise rating method needs to be remedied. This paper proposes an improved technique of quantifying the aspiration noise in the framework of direct waveform measurement.

First, jitter and shimmer of band-pass filtered waveforms around the F3 region were measured. Noise ratings, in which raters observe irregularity of the waveforms, were found to be highly correlated with aspiration noise. Therefore, it is natural to assume that either jitter, shimmer, or both, which are ratios of the irregularity of frequency and amplitude, also reflect aspiration noise. Second, in order to consider the interference of harmonics on the waveforms extracted above, the original waveform's jitter and shimmer were measured as references.

Finally, measurements of jitter and shimmer were employed for the comparison of the judgement of politeness and breathiness, as the latter has been suggested to express politeness and to show care. Listeners showed that there is a significant difference of shimmer around the F3 region, between utterances directed to people of both superior and inferior status.

From this result, it is reasonable to say that the shimmer in the F3 region is a possible cue when judging politeness.

1. Introduction

The vocal paralinguistic contribution to politeness has been studied by many researchers. However there are parts of this area still to be explored.

Brown and Levinson [4] suggested that the usage of paralinguistics seems to share a number of universal characteristics amongst cultures and language systems, but the strategies for expressing politeness are affected by the social and cultural background of the speakers. They also suggested that sustained high pitch may imply a self-humbling attitude and thus deference, and Ohala [13] also supported this concept of "frequency code", as a near-universal pattern.

However, in Japanese society, high F0 may be more associated with femininity than with politeness. Comparative studies of male and female speakers' speech by Ohara [14] revealed the tendency of expressing politeness with higher F0 by female speakers, which partially supports the use of frequency code. Interestingly, a contradictory tendency was reported for male speakers, who avoided the use of high-pitch. Analysing Japanese male speakers' politeness in speech may therefore reveal other vocal characteristics and strategies of expressing politeness.

In this study, the contribution of voice quality was the focus of interest. Up to now, breathiness was the target of observation, as Laver [12] associated this with familiarity or intimacy, that is, positive politeness. Since Laver [12] however, there have been several quantitative studies of voice quality. For example, Gobl [5] examined the correlates of four different voice qualities by extracting glottal parameters from the inverse-filtered waveform and measuring parameters on the glottal spectrum, according to Laver's taxonomical view. Gobl found that larger first-formant bandwidth (BW1) and steeper spectral tilt are likely to represent breathiness. In a recent study, Hanson [6] [7] [8] introduced an alternative to inverse filtering, which may filter out important characteristics of breathiness in the high frequency region, according to Klatt and Klatt [11]. Since these characteristics are likely to appear in the high frequency region, Hanson suggested the following method to measure the open quotient (OQ), first formant bandwidth (BW1), and spectral tilt. A steep spectral tilt is most observable between the lower frequency region where F0 lies and the higher frequency region, where the third formant peak lies, so the relative amplitude of the first harmonic and the third formant peak (H1-A3) is measured as a parameter of spectral tilt. Hanson also employed a noise rating technique to estimate aspiration noise, which is suggested by Klatt and Klatt [11]. In this method, the amount of aspiration noise in relation to the periodic component is estimated subjectively by examining the bandpass-filtered waveform in the third formant region. Klatt's group and Hanson both confirmed this method as valid since it showed good correlation between raters. Measuring BW1 on its own can be a predictor of either breathiness, whispery voice, or nasality, so an observation of above two trends in terms of spectral tilts and noise ratings in high-frequency region is therefore necessary. Regarding the above, our previous reports [10] demonstrated the following trends: 1) positive correlation overall (across speakers) between BW1 and politeness, 2) negative correlation between BW1 and politeness, when a speaker has narrow BW1 in modal voice, 3) positive correlation between aspiration noise estimated from noise ratings and politeness, when a speaker has wide

BW1 in modal voice, 4) negative correlation between spectral tilt and politeness, when a speaker has narrow BW1, 5) overall positive correlations between spectral tilt and noise rating. However, these two correlates, spectral tilt and noise ratings, are not free from problems as a way to measure breathiness. First, spectral tilt is not free of the effect of aspiration noise when a speaker has very breathy voice. Second, noise rating technique depends on subjective observations by raters. Thus it is subject to individual perception of visual irregularity. Therefore, establishing some objective method to quantify aspiration noise is necessary.

In this paper, we suggest the measurement of jitter and shimmer around the third formant frequency region as an alternative method to this noise rating technique. It is expected to serve as objective measures of aspiration noise.

2. Speech Data Collection

In order to collect data containing politeness and role information in a controlled setting, we recorded speakers performing the Map Task [1] [2]. The benefits of using the Map Task include the following. First, the formality can be maintained in a dialogue between participants. Since the role of the participants can alternate, the effects of relative status change would produce different suprasegmental features, which can be studied [9]. Second, the Map Task helps participants to concentrate on a specific task, and enables us to restrict the vocabulary and intentions of speakers, and thus to compare the suprasegmentals of lexically similar utterances.

Seven male Tokyo dialect speakers were recruited for this recording. In order to be assured that the participants were familiar with each other and had a similar background, but at the same time to control the status relationship, the participants were members of a single research group. Higher and lower status persons alternated in taking the role of an Instruction giver or an Instruction follower.

All materials were digitally recorded (16bit, sampling frequency = 48kHz, stereo) on DAT with a close-talking microphone and one DAT channel per participant. Down-sample rate for further analysis was 16kHz. Five speakers out of seven satisfied following conditions: 1) that they should talk to both higher status and lower status partners, 2) that they are native speakers of Tokyo dialect brought up in the Tokyo dialect area. Target utterances from these five speakers were extracted successfully, and were used for further analysis and experiment. In order to control phonological environment, such as pitch pattern and preceding/following consonants, vowel segments (/a/) of the word /hidari/ were extracted from natural utterances produced by five male native speakers of Tokyo dialect. Since the meaning of this word is “left”, the participants often produced this to instruct direction.

3. Noise Rating

As suggested by Klatt et al., [11] and Hanson [6] [7] [8], a noise rating test was conducted. In this method, the vowels were bandpass-filtered around the third formant using a filter with a bandwidth of 400 Hz. The bandpass-filtered waveforms corresponding to the speech segments used in the previously described measures were given ratings for noise, as described in the Introduction. These judgements were made independently by three students studying linguistics/music who took acoustics class for more than two terms. The raters were asked to estimate the amount of noise on a scale from 0 to 10, where 0 means there is essentially no evidence of noise interference and 10 means that there is little evidence of periodicity.

4. Jitter and Shimmer Measurement

In this study, there is a new attempt of measuring aspiration noise quantitatively. A method of jitter and shimmer measurement of the waveform from high-frequency region was used for examining aspiration noise. Hanson [6] mentioned that jitter and shimmer represent the irregularity of vocal fold vibration, whereas aspiration noise arises from posterior chink. However, if we extract the component of a waveform bandpass-filtered around the F3 region only, do a visual regularity check for noise ratings and use the extracted tokens for measurement of jitter and shimmer, the results of the measurement are likely to reflect quantitatively the irregularity of the waveform from aspiration noise. We can check if jitter and shimmer affect to the visual irregularity check of waveform, by computing correlations between jitter, shimmer and Nw. Therefore, it seems reasonable to measure jitter and shimmer from the band-pass filtered waveform, to find out if jitter and shimmer could be acoustic cues for aspiration noise and breathiness. For the measurement of jitter and shimmer, new functions of Praat (version 4.1.9) were employed for all the tokens band-pass filtered around the F3 region with the bandwidth of 400Hz that were used for noise ratings. Jitter was measured as follows. The Relative Average Perturbation (rap), which is the average absolute difference between a fundamental period and the average of it and its two neighbour periods, was divided by the average period. Shimmer was measured as follows. The three-point Amplitude Perturbation Quotient (apq3), which is the average absolute difference between the amplitude of a fundamental period and the average of the amplitudes of its neighbour periods, was divided by the average amplitude. For pathological examination, apq11 (the eleven-point APQ) is likely to be used for measurement of shimmer. However, the purpose of this study is to observe natural spontaneous speech, and getting eleven stable cycles of target vowel is extremely difficult. Therefore apq3 was employed as an available option. Additionally, jitter and shimmer of original waveform were measured in a similar way to each other. Jitter and shimmer of F3 were divided by jitter and shimmer of original waveform, and then taken to the log-scale. If the production of formality involves the tension of vibrating part of vocal folds rather than breathiness from posterior chink, jitter and shimmer of the original waveform will reveal the correlation between voice and politeness.

5. Perception Tests of Politeness

The utterances from which these vowels were taken were then presented to native Tokyo Japanese listeners. They were asked to rate politeness-related features of the utterances, in a way that aimed to distinguish the perception of formality from the perception of positive politeness. To deal with this problem, it was necessary to conduct two experiments concerning the perception of formality signalled by Keigo. Firstly, a formality judgment experiment was carried out based on the text level presentation. To measure formality expressed with Keigo in the text level, the Magnitude Estimation method (ME) was employed. This method was originally developed for psychophysics, although it has since been used by Bard et al. [3] in psycholinguistics. Secondly,

subjects were asked to make a forced-choice judgment of the relative status of speaker and addressee, based on recordings of the map task, as discussed below. This judgment gives a measure of the influence of positive politeness signalled paralinguistically.

To minimise the effects on prosodic features from dialect and regional cultural backgrounds, only native speakers of Tokyo dialect were recruited to participate in this experiment. A total of 22 people participated as subjects. A group of 15 subjects were students without any work experience, and the other 7 subjects have work experience.

In this experiment, the aim was to compare listeners' reactions with acoustic differences, so the utterances, which include the contextually similar vowel /a/, were carefully chosen as materials from previous speech collection, taking care to avoid semantic and contextual influences. Utterances which were judged disfluent were excluded. Utterances were chosen, which were produced while addressing both higher and lower status participants. Utterances, which are in a phonetically similar environment, are suitable for observing voice quality and assimilation. A set of 160 tokens, consisting of 32 tokens per each of five speakers, was extracted as test stimuli.

Subjects were tested individually. In the Keigo formality judgment, they were first given a practice model to follow, and then presented a set of stimuli, sentence by sentence from the collection of materials, and were required to estimate the magnitude of formality of each stimulus. Each input field was displayed on the PC, together with a stimulus sentence, and the subjects were asked to give their estimated score. A set of stimuli consisted of 16 phrases includes /hidari/ from each speaker, randomly mixed with 16 dummy utterances. Each session started with a speaker's modulus followed by a set of these 32 stimuli. The subjects were presented with one set for each of the five speakers. Thus a total of 160 stimuli were presented. Since ME data follow a power scale, they were processed logarithmically and normalised by subject, so that raw scores were converted to a linear scale, ranging between 0 and 1 for each subject. Then a mean value of each stimulus among the subjects was computed as Keigo-formality score (F score).

In the forced-choice experiment, the subjects were presented each utterance in speech and were asked to judge whether the addressee was of higher status, which was scored as 1, or lower status, which was scored as 0. The agreement rate between the subjects for each stimulus was also computed as the S score. Because both this agreement rate and the formality judgment score take a value between 0 and 1, it allows us to compare the two scores with each other.

6. Result and Discussion

To investigate which acoustic correlates are influential in judging relative status, the correlations between the S score and acoustic parameters, including Nw (noise ratings) were computed, using a one-way ANOVA. The means and standard deviation were also computed. The whole set of the S scores was computed in groups as follows.

Group 1: S score < 0.4
(highly agreed as an utterance directed at an inferior).
Group 2: S score > 0.6
(highly agreed as an utterance directed at a superior).
Group 3: (S score - F score) < -0.15
(speech contributed to decreasing the degree of formality)
Group 4: (S score - F score) > 0.15
(speech contributed to increasing the degree of formality)

Table 1: *The mean and standard deviation of jitter ratio, and shimmer ratio, shimmer around F3, and Nw, according to the high agreement of the S score (squared by bold line: significance level (p < 0.01, by ANOVA test.)*

ID	Group	Nw	shimmer (adp3) F3	jitter (rap) F3/ref	shimmer (adp3) F3/ref
M1	1 (N=6)	6.889 (0.455)	2.309 (0.190)	-0.973 (0.102)	-0.229 (0.268)
	2 (N=1)	4.667 (.)	2.488 (.)	-1.122 (.)	-0.382 (.)
M2	1 (N=9)	6.963 (1.821)	3.773 (1.845)	-1.000 (0.273)	0.040 (0.085)
	2 (N=4)	7.583 (0.995)	3.356 (1.209)	-1.168 (0.266)	-0.187 (0.088)
M3	1 (N=12)	6.806 (1.586)	4.210 (2.807)	-0.901 (0.185)	-0.008 (0.121)
	2 (N=1)	3.000 (.)	0.399 (.)	-0.368 (.)	0.642 (.)
M4	1 (N=5)	5.867 (1.709)	4.888 (2.521)	-1.037 (0.303)	-0.938 (0.310)
	2 (N=10)	8.400 (1.173)	3.818 (0.823)	-0.938 (0.310)	-0.292 (0.355)
M5	1 (N=6)	5.800 (1.807)	3.736 (1.413)	-1.083 (0.243)	-0.0127 (0.260)
	2 (N=1)	6.333 (1.333)	5.203 (3.039)	-1.184 (0.337)	0.114 (0.210)

Firstly, Group 1 and Group 2 were compared, using a one-way ANOVA (Table 1). For Speaker M1, no acoustic correlates showed a significant difference between the groups. For Speaker M2, the shimmer ratio showed a significant difference between the groups ($F(2.13)=7.13$, ($p<0.01$)). For Speaker M3, the jitter ratio, and shimmer ratio showed significant differences between the groups (the jitter ratio: $F(2.15)=5.00$, ($p<0.03$); and the shimmer ratio: $F(2.15)=11.37$, ($p<0.01$);). However, it should be considered that this speaker had only one token in Group 2.) For Speaker M4, the shimmer in F3 region and Nw showed significant differences between the groups. (the F3 shimmer: $F(2.16)=5.97$, ($p<0.02$); Nw: $F(2.16)=5.64$, ($p<0.02$)). For Speaker M5, a slight but significant difference between the groups was found at the F3 shimmer ($F(2,13)=3.37$, ($p=0.06$)). A total of four speakers showed that the shimmer ratio F3/ref or the F3 shimmer was significant in judging the relative status of the addressee. From this result, we may assume that the shimmer observed in F3 region may be involved in judging the relative status of addressee. If so, the irregularity of the waveform amplitude in the high frequency region affects the perception of politeness.

Secondly, Group 3 and Group 4 were compared, using a one-way ANOVA. For the speaker M1 and M2, the results were almost the same as the previous comparison between Group 1 and Group 2. For Speaker M1, no acoustic correlates showed a significant difference between the groups. For

Speaker M2, the shimmer ratio showed a significant difference between the groups ($F(2.13)=3.88$, ($p<0.05$)). However, for Speaker M3, the tendency is completely different. No significant difference between the acoustic correlates of the groups was found. For Speaker M4, only Nw showed a significant difference between the groups. (Nw: $F(2.16)=6.59$, ($p<0.01$)). For Speaker M5, no significant difference between the groups was found. Therefore, it is difficult to say which acoustic correlate had an influence on increasing or decreasing the perceived status. We cannot see what kind of voice quality has the most effect on changing the impression of politeness.

Table 2: *The mean and standard deviation of Nw and the shimmer ratio, according to the high contribution to the S score (squared by bold line: significance level ($p < 0.01$, by ANOVA test.)*

speaker	Group	Nw Mean(s.d.)	Shimmer(adp3) F3/ref Mean(s.d.)
M1	3 (N=7)	6.333(1.317)	-0.271(0.159)
	4 (N=1)	8.000(.)	0.173(.)
M2	3 (N=6)	6.890(1.559)	0.062(0.090)
	4 (N=5)	8.133(1.121)	-0.109(0.163)
M3	3 (N=8)	6.375(1.174)	0.015(0.088)
	4 (N=4)	6.083(2.587)	0.165(0.354)
M4	3 (N=4)	5.250(1.167)	-0.323(0.403)
	4 (N=7)	7.429(1.978)	-0.113(0.222)
M5	3 (N=8)	6.167(1.369)	-0.0246(0.290)
	4 (N=3)	6.444(0.962)	0.028(0.112)

From the results of the perception experiments, it is possible to say that the listeners used vocal paralinguistic features for judging the relative social relationship between the speakers. Shimmer in the F3 region, as a quantitative component for measuring aspiration noise, was correlated with the listeners' assured responses. Therefore, we can still assume that breathiness is likely to have a role in perceiving social relationships.

However, the directions of the associations between F3 shimmer and judged status differ from speaker to speaker. Therefore, we need to explore how listeners use the acoustic correlates case by case. For this purpose, voice data from a large number of speakers should be studied.

7. Acknowledgements

This research is based on my PhD work at the University of Edinburgh. I would like to thank my supervisors, Prof. Ladd and Dr. Turk for their helpful comments. I would like to thank Prof. Tsuchiya and Dr. Horiuchi (Chiba University), for allowing me access to the materials of their Map Task corpus, and giving many helpful technical advices before the recordings. I am also grateful for Prof. K. Hirose (the University of Tokyo) to allow me access to his laboratory's facilities and staff through these experiments. However, any mistakes that remain are my own.

8. References

[1] Anderson, A.H. et al.; 1991. "The HCRC Map Task Corpus", *Language and Speech*, 34, 351-366.

[2] Aono, M. et al.; 1994. The Japanesse Map Task Corpus: An interim report (in Japanese). *Spoken language understanding and discourse processing, Japanese Society for Artificial Intelligence,* SIG-SLUD-9402, 25-30.

[3] Bard, E.G.; Robertson, D.; and Sorace, A.; 1996. "Magnitude Estimation of Linguistic Acceptability" *Language*, 72, 32-68.

[4] Brown, P.; S. Levinson, S.; 1987. *Politeness: Some universals in language usage*, Cambridge University Press, Cambridge.

[5] Gobl C., 1989. "A preliminary study of acoustic voice quality correlates", *STL-QPSR*, 4, . 9–22, Stockholm.

[6] Hanson, H.M.; 1995. *Glottal characteristics of female speakers.* PhD thesis, Harvard University, Cambridge, MA, 1995.

[7] Hanson, H.M.; 1997. "Glottal characteristics of female speakers: Acoustic Correlates" *J. Acousitc Society of America*, 101(1), 466-481.

[8] Hanson, H.M.; and Chuang, E.S.; 1999. "Glottal characteristics of male speakers: Acoustic correlates and comparison with female data" *J. Acousitc Society of America*, 106(2), 1064-1077.

[9] Ito, M.; 2002. "Japanese Politeness and Suprasegmentals: A Study based on Natural Speech Materials", B. Bel & I. Marlien (eds.), *Proc. of the Speech Prosody 2002 conference*, Aix-en-Provence, 415-418.

[10] Ito, M.; 2003. "The Contribution of Voice Quality to Politeness of Japanese", *Proceedings of VOQUAL03*, Geneva.

[11] Klatt, D.; and Klatt, L.; 1990. "Analysis, synthesis, and perception of voice quality variations among female and male talkers" *J. Acousitc Society of America*, 87, 820-857.

[12] Laver, J.; 1980. *The Phonetic Description of Voice quality*, Cambridge University Press.

[13] Ohala, J. J.; 1996. "Ethological Theory and the Expression of Emotion in the Voice", *Proc. ICSLP*, Philadelphia, 3, 1812-1815.

[14] Ohara Y.; 2001. "Finding one's voice in Japanese: A study of the pitch levels of L2 users", in Pavlenko, A., Brackledge, A., Piller, I. and Teutsch-Dwyer, M. (eds.), *Multilingualism, Second Language Learning, and Gender*., New York: Mouton de Gruyter.