Voice Quality: the 4th Prosodic Dimension

Nick Campbell[†] **& Parham Mokhtari**[‡]

† ATR Human Information Science Laboratories

‡ JST/CREST Expressive Speech Processing

E-mail: {nick, parham}@atr.co.jp

ABSTRACT

This paper presents data from an analysis of a large conversational-speech corpus, showing evidence that voice quality, as measured on a continuum from pressed to breathy using a normalized amplitude quotient (NAQ), is varied consistently, and in much the same way as, but independently of, fundamental frequency, to signal paralinguistic information. We show that interlocutor, speaking-style, and speech-act all have significant interactions with NAQ, and argue that voice-quality should be considered as the 4th prosodic parameter, along with pitch, power, and duration.

1. INTRODUCTION

The history of prosodic research closely follows the history of speech processing technology. Pitch information (or more accurately, fundamental frequency of the voice, hereafter "f0") was relatively easy to extract from the speech signal, as was amplitude (or RMS signal power), but it was not until the nineteen-eighties that speech databases were large enough, computers fast enough, and labelling software robust enough, for speech timing or duration studies to appear in the academic literature.

The 21st century brings new advances in speech processing technology, amongst them the robust extraction of voice-quality information. Based on an algorithm proposed by Alku [1], and extended by the second author [2] to work on fluent conversational speech , we are now able to extract a measure of "breathiness" in everyday speech.

This paper proposes that, following the advances in signal processing, we should now include voice-quality as the 4th prosodic dimension, and shows how breathiness and tension in the voice are used to convey social and paralinguistic information, in much the same way as do variations in f0, amplitude, and local speaking rate.

2. THE AMPLITUDE QUOTIENT

The mode of glottal phonation (or "voice-quality") can be measured from an estimate of the glottal speech waveform derivative (a result of inverse filtering of the speech using time-varying optimised formants [2] to remove vocal tract influences) by calculating the ratio of the largest peak-to-peak amplitude and the largest amplitude of the cycle-to-cycle minimum derivative. This is the Amplitude Quotient (AQ) as described by Alku in [1]. In its raw form, it is weakly correlated with the fundamental period of the speech waveform, but this can be greatly reduced by normalisation based on f0, giving a normalised measure of the Amplitude Quotient (henceforth "NAQ") [3].

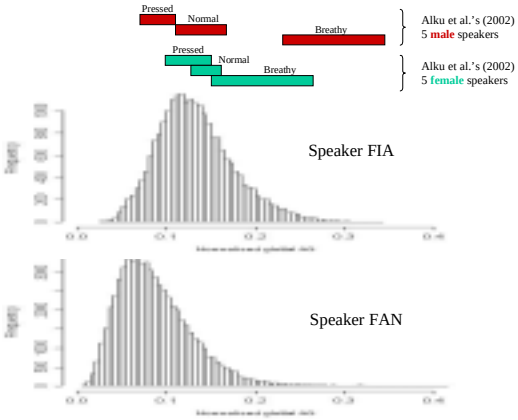

Figure 1. NAQ plotted for 2 speakers of our corpus.

Figure 1 shows histograms of NAQ measured from the speech of two Japanese females (FIA & FAN), comparing them with the measures reported in [3]. We can see that there is some individual variation, but that the overall shape of the distributions is similar, and that they both fit into the range of "pressed", "normal", and "breathy" reported in [3] . The skew observed in the data of speaker FAN can be accounted for by the predominance of a more casual (pressed) speaking style as explained below.

We will show below that this variation is not random, and that it can be best explained by examining correlations with paralinguistic features of the speech, such as "interlocutor relations", "speaker intention", and "speaking-style", and that, as such, it should be considered a prosodic parameter.

3. CONVERSATIONAL SPEECH DATA

The JST/CREST Expressive Speech Processing project is collecting a large amount (1000 hours) of fluent conversational speech in three languages (English, Chinese, and Japanese) using high-quality recordings of everyday social interactions [4].

A portion (about 250 hours) of this speech has been transcribed and a smaller portion (about 100 hours)

annotated for speaking style and speech-act features. Acoustic measures of the speech have been produced for an analysis of correlations between the perceptual and physical attributes [5].

We will concentrate in this paper on the data from one female Japanese speaker, who wore a small head-mounted, studio-quality microphone and recorded her day-to-day spoken interactions onto a MiniDisk [5]. Only the voice of the main speaker was analysed, but the speech of the interlocutors was sometimes also (albeit faintly) available to the labellers who produced the perceptual annotations.

The data consist of 13,604 utterances, being the subset of the speech for which we had satisfactory acoustic and perceptual labels. An "utterance" can be loosely defined as the shortest section of speech having no audible break, and perhaps best corresponds to an "intonational phrase". (By paying our transcribers per utterance unit, we encourage shortest meaningful units). They can vary in length from a single syllable to a thirty-five syllable stretch of speech.

The data were analysed using the public-domain "R" statistical software package from CRAN [6]. Feature sets consisting of labels for interlocutor ("who"), speaking-style ("how"), and speech-act ("what") were produced, and matched with measures of NAQ and f0 for an analysis of any correlations.

Interlocutors were grouped into the classes listed in Table 1. Speaking style labels were simplified and grouped for this study into "polite", "friendly", and "casual", for each of the classes "family", "friends", and "others", with an extra category for self-directed speech. There were 24 speech-act categories, of which we will focus on 5 here: "giving information", "exclamations", "requesting information", "muttering", and "requesting repeats".

Table 1. Counts of utterances grouped by interlocutor.

Child	Family	Friends	Others	Self
139	3623	9044	632	116

4. SPEECH PROSODY AND NAQ

Before normalisation, the AQ measure per syllable showed a correlation of r = -0.406 with f0. After normalisation (by NAQ = log(AQ) + log(f0)), the resulting NAQ showed a weak correlation of only r = 0.182 with f0. Figure 2 plots this relationship. We can see that there is a wide spread of NAQ around the middle of the f0 range, and will confirm below that the two function independently in all quadrants but that of the very lowest f0, and with significant correlations to the perceptual categories described above.

Figure 3 plots median NAQ and f0 for the five categories of interlocutor. We can see that NAQ is highest (i.e, the voice is breathiest) when addressing "others" (talking politely), and second highest when talking to children (softly). Self-directed speech shows the lowest values for NAQ.

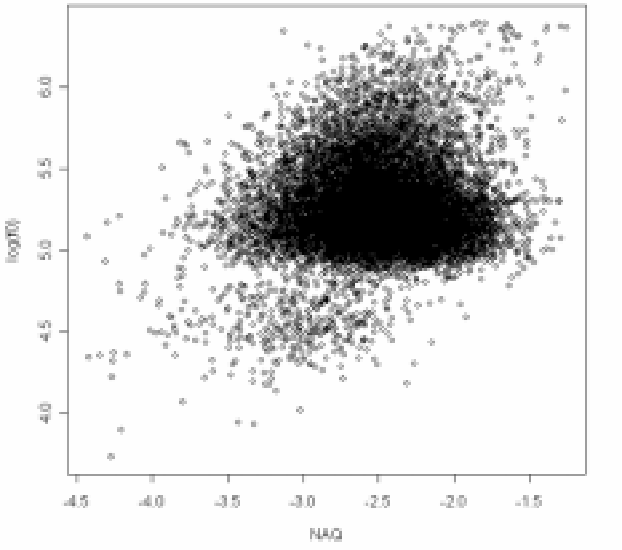

Figure 2. Plot of log f0 and NAQ for Speaker FAN.

Speech with family members exhibits a higher degree of breathiness (high NAQ) than that with friends. F0 is highest for child-directed speech, and lowest for speech with family members (excluding children). It is clear from the figure that f0 and breathiness are being controlled independently (though whether consciously or not, and through what mechanisms, remains a matter for separate investigation) for each class of interlocutor.

Table 2 shows the results of a pairwise t-test for the NAQ data plotted in Figure 3. The false discovery rate (i.e., the expected proportion of false discoveries amongst the rejected hypotheses) was controlled by the FDR method [7]. The repeated t-tests confirm all but the child-directed (n=139) voice-quality differences to be significant.

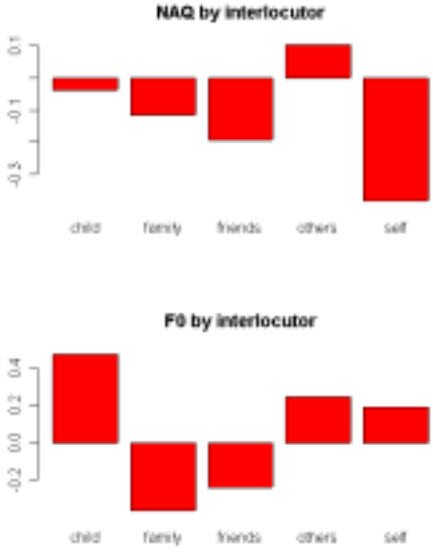

Figure 3. Median values of NAQ and F0 plotted for interlocutor. The data are (z-score) scaled, so values are in SD units. 0 represents the mean of the distribution.

Table 2. Pairwise t-tests showing adjusted p-values. All differences of NAQ are significant except those for child-directed speech, which has the smallest number of utterances and a large variation (see Fig 8 below).

	child	family	friends	Others
family	0.58 (ns)	-	-	-
friends	0.10 (ns)	2.7e-05	-	-
others	0.16 (ns)	0.00042	1.8e-08	-
self	0.00143	0.00042	0.00656	2.0e-06

Figure 4 plots the values for "family" speech in more detail. It reveals some very interesting tendencies. We can see that speech directed to her daughter (a one-year-old) exhibits both the highest F0 and the highest breathiness. Family members can be ordered according to breathiness as follows: daughter > sister's son > father > mother = older sister > aunt > husband. Thus, it seems that this ordering reflects the degree of "care" taken in the speech to each family member. The labelers, too, confirm that this accords with their subjective experience of listening to the speech.

Analysis of speech-act and speaking style features also supports this interpretation. We can see from Figure 5 (Manner) that the four categories show only three levels of difference for f0, but four levels for NAQ. Class "0" (self-directed speech) exhibits the lowest NAQ values, whereas class "a" (*careful speech*) exhibits the highest.

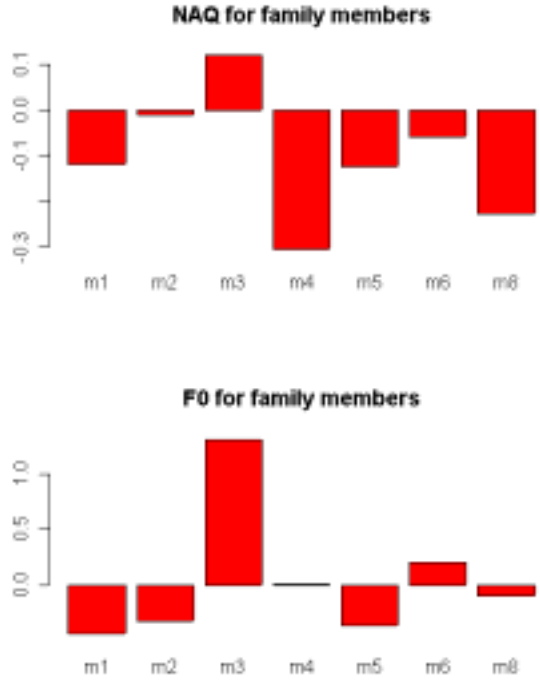

Figure 4. Median values of NAQ and F0 for family members. m1: mother, m2: father, m3: daughter, m4: husband, m5: older sister, m6: sister's son, m8: aunt.

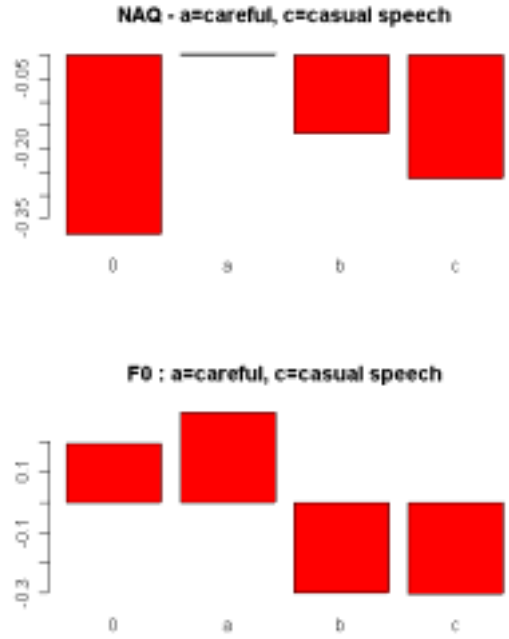

Figure 5. Manner: NAQ and F0 for 4 speaking styles.

Class "c" (*casual*) has lower NAQ than class "b" (*friendly*). These data are consistent with the view that NAQ corresponds with "degree of care" in the speech. F0 values for these classes do not show the same ordering, but careful speech yields a higher f0 than friendly or casual speaking styles. Figure 6 shows the same values, factored by class of interlocutor. Here, "f" stands for friend, "m" for family, and "t" for others. It is interesting that NAQ for careful speech is high for friends (fa), while there is no difference between friendly (fb) and casual (fc), whereas for family members the difference is reversed: there is no difference between careful (ma) and friendly (mb) utterances, but significantly lower NAQ for casual ones (mc). For speech with others, there were no casual utterances but careful and friendly ones showed the expected NAQ differences.

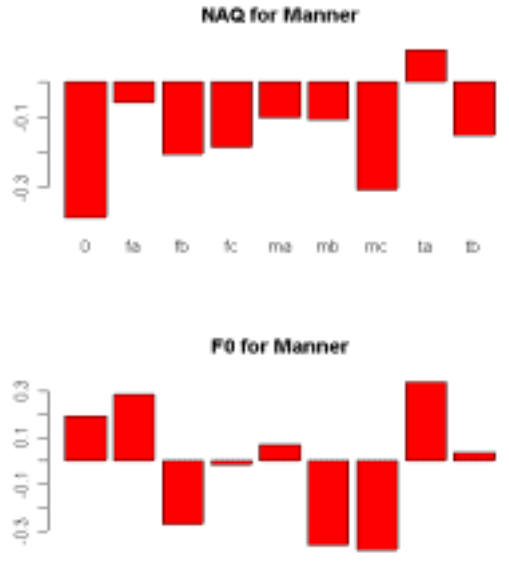

Figure 6. Median NAQ and F0 for style by interlocutor.

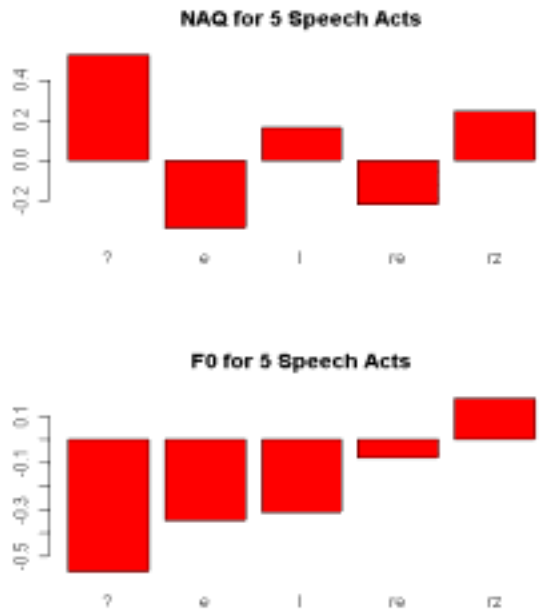

Figure 7. NAQ and f0 for 5 speech-act types.

Finally, we turn to speech-act differences. From the above, we might expect careful acts to show higher NAQ than less socially expensive ones. Figure 7 confirms this to be the case; the 5 categories presented here are muttering ("?"), interjections ("l"), giving information ("e"), requesting information ("re"), and requesting repeats ("rz"). We find that *giving information* has significantly lower NAQ values (t=3.2805, df=1453.04, p=0.001061) than *requesting information*, and that *requesting repeats* (requiring most effort on the part of the interlocutor) has the highest NAQ. The figure confirms auditory impressions that *muttering* may be considered a separate category, with a distinctively lower f0 and a breathy (high NAQ) voice quality.

5. CONCLUSION

This paper has shown that voice-quality, measured by Normalised Amplitude Quotient (NAQ), has significant correlations with interlocutor, speaking-style, and speech-act. It varies consistently with degree of "care" in the speech, and varies independently of fundamental frequency. We therefore conclude that voice-quality should be considered a prosodic characteristic, along with f0, duration, and amplitude, and is controlled in the speech production to signal paralinguistic differences in meaning.

We note that this discovery is a result of advances in speech signal processing technology that enable the robust measurement of breathiness in fluent conversational speech. We note also that the figures presented above show only the median values for NAQ and F0, and that a better understanding of the distribution and overlap of these parameters may be gained from boxplots of the data. Figure 8 plots the same data as Figure 3. We can see that although the median values reveal the trends (confirmed by pairwise t-tests), the data contain utterances of varying voice quality, and a finer classification of effects is needed.

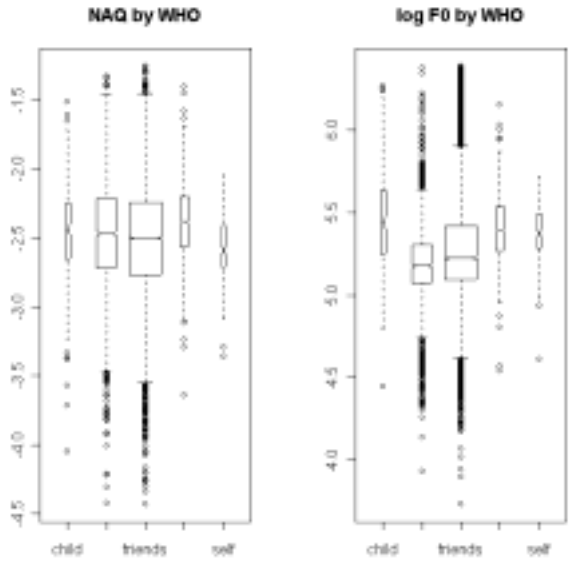

Figure 8. Boxplots of NAQ and f0 interlocutor data.

Acknowledgements

We are grateful to the Japan Science & Technology Corporation for support, and to Minako Kimura and the ESP labelers for their continuing help and advice.

REFERENCES

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

[2] P. Mokhtari, and N. Campbell, "Automatic Measurement of Pressed/Breathy Phonation at Acoustic Centres of Reliability in Continuous Speech" in Trans IEICE Special Issue on Speech Information Processing, March 2003.

[3] P. Alku, T. Bäckström, and E. Vilkman, "Normalized amplitude quotient for parametrization of the glottal flow", J. Acoust. Soc. Am., vol.112, no.2, pp.701-710, 2002

[4] N. Campbell, "Recording Techniques for capturing natural everyday speech", in Proc Language Resources and Evaluation Conference (LREC-2002), Las Palmas, Spain, 2002.

[5] N. Campbell, and P. Mokhtari, "DAT vs. Minidisc: Is MD recording quality good enough for prosodic analysis?", 1-P-27, in Proc Acoustical Society of Japan Spring Mtg., 2002.

[6] Comprehensive R Archive Network: cran.r-project.org

[7] Y. Benjamini, and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing". Journal of the Royal Statistical Society Series B, 57, pp. 289-300, 1995.