CHANGES IN VOICE QUALITY DUE TO SOCIAL CONDITIONS

Nick Campbell

National Institute of Information and Comunications Technology
& ATR Spoken Language Communication Research Labs,
Keihanna Science City, Kyoto 619-0288, Japan
nick@nict.go.jp, nick@atr.jp

ABSTRACT

This paper describes how acoustic features of the voice vary according to social relationships between speakers, and proposes that voice quality is an important aspect of prosodic information which serves to carry this separate strand of affect-related information, in parallel with variation according to the linguistic information in a spoken discourse.

Keywords: Voice quality, Speech communication, Interpersonal stance, Conversational speech, Corpus-based analysis, Statistical processing

1. INTRODUCTION

The acoustic characteristics of speech are usually modelled as a sequence of source, vocal tract filter, and radiation characteristics [1]. Of these, glottal closure has an important influence on the quality of the generated speech signal and it is associated with perceived breathiness [2]. This aspect of voice quality has been studied extensively [3] but is not typically included as a prosodic parameter in discussions of speech prosody. Campbell & Mokhtari recently referred to voice quality as the fourth prosodic dimension [4] showing that it varies consistently according to speaker intention and speaker-listener relationships. This paper provides support for that claim and shows that speakers do consistently vary their vocal settings according to social factors.

Whereas there has been considerable research into the linguistic nature of prosodic variation, there has to our knowledge been less research into its social aspects. The work presented here is based upon an analysis of a large corpus of high-quality conversational speech recordings, with a statistical analysis of several acoustic features factored according to interlocutor and familiarity.

2. DATA

The speech data were recorded over a period of several months, with paid volunteers coming to an office building in a large city in Western Japan once a week to talk with specific partners in a separate part of the same building over an office telephone. While talking, they wore a head-mounted Sennheiser HMD-410 close-talking super-cardiod microphone and recorded their speech directly to DAT (digital audio tape) at a sampling rate of 48kHz. They did not see their partners or socialise with them outside of the recording sessions. Partner combinations were controlled for sex, age, and familiarity, and all recordings were transcribed and time-aligned for subsequent analysis. Recordings continued for a maximum of ten sessions between each pair. Each conversation lasted for a period of thirty minutes.

In all, ten people took part as speakers in these recordings, five male and five female. Six were Japanese, two Chinese, and two native speakers of American English. All conversations took place in Japanese. There were no constraints on the content of the conversations other than that they should occupy the full thirty-minute time slot. Partners were initially strangers to each other, but became friends over the period of the recordings. The conversations between the three pairs of Japanese speakers form the main part of this corpus, and the conversations with non-native speakers form a sub-part. A further sub-part consists of a series of conversations between two of the Japanese speakers and their own family members, using the same telephone setup. The speech of the family members, who were not present in the building, was not recorded. The conversations were thus balanced for familiarity of the partners, ranging from highly familiar, through unfamiliar (but of the same cultural background) to unfamiliar and from a different cultural background. The non-native speakers were competent in Japanese, but not at a level approaching native-speaker fluency.

The speech data were transferred to a computer and segmented into separate files, each containing a single utterance. The definition of a speech utterance is not simple; transcribers were asked to segment the speech as finely as they could without splitting a coherent "meaning unit". This resulted in many short single-word utterances (backchannels such as "yes",

"uhuh", etc.) and some considerably longer utterances of up to fifty syllables. The majority of utterance units can be said to correspond to a minor intonation unit [5].

The transcriptions were made in Japanese, which has a phonetic orthography, and the mappings between the transcription and the speech are phonemic. No further fine phonetic analysis was performed, but length diacritics were use to indicate moraic lengthening, as is provided for in the kana orthography. Special symbols were included in the text to indicate non-speech noises such as lip-smacks, sucking-in of breath, coughing, laughter, etc. The analysis focusses on the speech of the Japanese native speakers and observes the way it changes according to relationship with the interlocutor and progression of their familiarity throughout the series of conversations.

3. RESULTS

As we are principally concerned with the relation between voice quality and the expression of affect, we limit our analysis to those shorter utterances that occurred more than a threshold of 100 times each in the conversations. The resulting 67,792 utterances include many backchannel utterances, used to indicate comprehension, attention, understanding, and interest, or to encourage the speaker to continue talking. They are complemented by laughter, greetings and phatic idiom, such as "Great game last night!", or expressions about the weather, etc.

The corresponding speech files were processed to obtain a set of acoustic features for each utterance. The parameters included pitch, power, duration, and spectral shape. Pitch was described by the mean, maximum, minimum, location of the peak in the utterance, and degree of voicing throughout the utterance. Power was described by the mean, maximum, minimum, and location of the peak in the utterance. Duration of the whole utterance was expressed as a log value, and a simple estimate of speaking rate was made by dividing the duration by the number of moraic units in the transcription. Spectral shape was described by the location and energy of the first two harmonics, the amplitude of the third formant, and the difference in energy between the first harmonic and the third formant (h1-a3, proposed by Hansen as the best measure for describing breathiness in her study of the voice quality of female speakers [6]). All these measures were produced automatically using the Tcl/Tk "Snack" audio processing library [7]. Thus for each common utterance in the conversations, we have a transcription and a vector of values corresponding to its acoustic characteristics.

Table 1: Results of a principal component analysis. SD indicates standard deviation of the component, PoV portion of the overall variance that it accounts for, and CP the cumulative portion accounted for by current and previous components.

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
SD	1.98	1.47	1.30	1.20	1.08	0.98	0.86	0.83
PoV	0.28	0.15	0.12	0.10	0.08	0.06	0.05	0.05
CP	0.28	0.43	0.56	0.66	0.74	0.81	0.87	0.92

	PC9	PC10	PC11	PC12	PC13	PC14
SD	0.72	0.63	0.372	0.234	0.004	0.0003
PoV	0.03	0.02	0.009	0.003	0.000	0.0000
CP	0.95	0.98	0.996	1.000	1.000	1.0000

Table 2: Rotation of the fourteen acoustic features according to the first six principal components.

Rotation (all ten speakers):

	PC1	PC2	PC3	PC4	PC5	PC6
dn	−0.20	0.52	0.32	0.02	0.14	0.05
fmean	−0.42	0.00	−0.21	−0.21	−0.12	0.23
fmax	−0.41	0.09	−0.15	−0.18	−0.08	0.12
fmin	−0.29	−0.14	−0.22	−0.18	−0.14	0.51
fpct	−0.02	0.05	−0.19	−0.19	0.67	0.19
fvcd	−0.20	0.52	0.32	0.02	0.14	0.05
pmean	−0.26	−0.37	0.41	−0.08	0.06	−0.03
pmax	−0.30	−0.19	0.43	−0.08	0.10	−0.14
pmin	−0.09	−0.44	0.31	−0.01	0.10	0.06
ppct	0.04	−0.13	−0.22	−0.21	0.59	−0.20
h1h2	0.01	0.04	−0.02	−0.54	−0.25	−0.33
h1a3	0.38	0.03	0.23	−0.36	−0.03	0.34
h1	0.36	−0.03	0.23	−0.09	0.01	0.52
a3	−0.13	−0.13	−0.07	0.59	0.08	0.22

3.1. Principal Component Analysis

A principal component analysis [8] was performed to simplify analysis of the acoustic features, using the "princomp" function of the R statistical programming language [9]. PCA is "an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on" (Wikipedia).

The first three coordinates accounted for more than 50% of the variance in the acoustic data, and the first six components accounted for more than 80%. Table 1 shows results of the pca transform, and Table 2 shows the relationship of the acoustic features to each of the component vectors. We can see from Table 1 that pc1 accounts for 28% of the variance and that pc2 accounts for 15%, bringing the cumulative proportion to 43%. By pc3, we have accounted for 56% of the overall variance of the acoustic features by this rotation of the feature space.

Table 3: Some common utterances from speaker JFA to different interlocutors. All conversation participants except JFB were in group A (the third letter of the speaker identifier); M or F (the second letter) indicates male or female respectively; and J, C, or E (the first letter) indicates native language; Japanese, Chinese or English, respectively.

	CFA	CMA	EFA	EMA	JFB	JMA
a	88	79	40	51	4	0
a-	31	38	11	46	14	14
ano	141	133	80	45	29	46
ano-	69	80	128	117	75	147
e	33	22	17	11	9	1
e-	12	23	17	13	2	7
etto	16	25	13	6	1	0
fun	50	24	19	54	34	151
fu-n	12	33	28	31	13	24
hai	81	89	13	88	61	26
he-	33	42	16	30	1	4
nee	11	13	14	7	15	39
sono	15	8	8	10	2	5
sou	14	4	4	1	30	10
un	915	415	463	977	799	947
u-n	38	49	35	78	213	88
unun	50	6	14	28	51	108
ununun	19	1	5	8	27	73
laugh	201	174	350	228	0	0

Table 4: Showing rotation of the fourteen acoustic features according to the first six principal components for 4,516 utterances of the word "un" spoken by one female speaker. JFA.

Rotation (speaker JFA's 'un' data):

	PC1	PC2	PC3	PC4	PC5	PC6
dn	0.21	0.35	0.45	-0.09	0.23	0.10
fmean	-0.42	0.17	-0.04	-0.21	-0.16	0.23
fmax	-0.28	0.36	-0.08	-0.16	-0.35	-0.04
fmin	-0.36	-0.11	0.00	-0.12	-0.03	0.49
fpct	-0.01	0.05	-0.35	-0.22	0.61	0.29
fvcd	0.21	0.35	0.45	-0.09	0.23	0.10
pmean	-0.42	0.00	0.26	-0.10	0.21	-0.08
pmax	-0.38	0.15	0.19	-0.14	0.07	-0.22
pmin	-0.30	-0.17	0.11	0.05	0.29	-0.31
ppct	0.01	0.01	-0.40	-0.32	0.37	-0.33
h1h2	0.14	0.01	0.05	-0.54	-0.24	-0.47
h1a3	0.07	-0.49	0.24	-0.36	-0.03	0.15
h1	-0.06	-0.51	0.30	-0.06	0.05	-0.00
a3	-0.24	-0.01	0.08	0.52	0.14	-0.28

Table 2 shows that pc1 is best explained by a combination of four acoustic features; fmean .42, fmax .41, h1a3 .38, and h1 .36. Thus the first principal component correlates well with pitch and voice quality. The second correlates with duration and amount of voicing (.52 each with pmin third at .44), and the third principal component with power (pmax and pmean at .43 and .41 respectively, with duration coming in third place at .32). For reasons of space, values for the lower components are not shown.

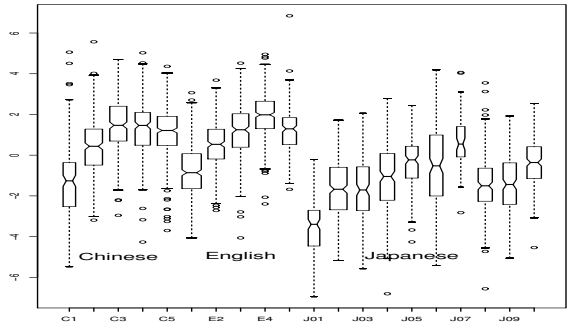

Figure 2: Development of JFA's pc1 across time, with 5 conversations each for the Chinese and English partners, and ten each for Japanese partners. All data for the single word "un" ('yes').

It accords well with our intuitions about speech prosody that fundamental frequency should occur as a strong variable influencing the first principal component, and it is no surprise that duration and power also appear as strong factors in the first three. It is interesting, however, that spectral shape should also appear so early as a contributing factor, and that the h1-a3 value appears in the first component lends support to Campbell & Mokhtari's claim that voice quality should also be considered as one of the controlled prosodic features in speech. We will see in the following sections that this difference is not just due to different speakers and different phonetic contexts but also to social factors.

3.2. Speaker/Partner Characteristics

Table 3 lists the most common utterances of one female Japanese speaker (JFA). "Ano" and "eto" are common hesitation markers in Japanese, similar to "umm" and "er" in English, and "fun" might be transcribed as "hmm" in English, "a" as "ah!". We see also that her use of language differs according to the nature of the interlocutor, with relatively little use of "a" with Japanese partners, and much more use of the polite hesitation marker "ano" in speech with Chinese partners, lengthening it to "ano-" when speaking with others. She tends to use "fun" ("hmm") more with the male Japanese partner, and "unun" (agreement) more with him too. Note that the apparent lack of laughs with the Japanese partners in this table is due to the fact that laughs were transcribed phonetically in the Japanese-Japanese conversations and are not so simple to count (but see paper at satellite workshop).

Table 4 shows the equivalent pca results for 4,516 instances of the word "un" in the speech of JFA, for comparison with the data for all ten speakers shown in Table 2. We see that h1-a3 appears strongly (.49) in the second pc component. Figure 1 plots differ-

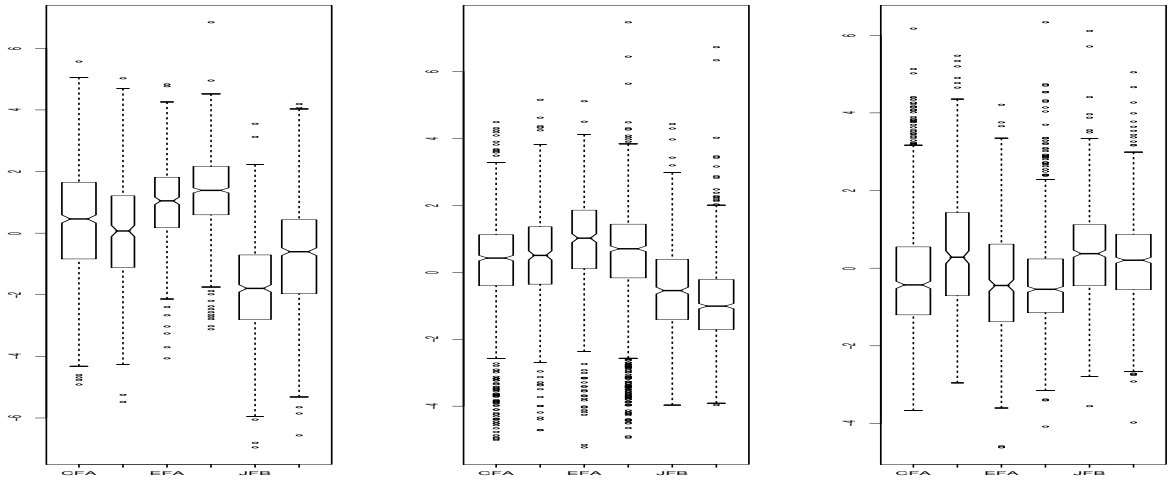

Figure 1: Showing how the principal component coefficients vary according to partner. The left plot shows pc1 for the word "un" from speaker JFA, the middle plot pc2, and the third plot pc3. Partners are CFA CMA, EFA EMA, JFB JMA respectively.

ences in the values of pc1-3 for each partner. We can see a clear and significant difference in vocal settings for each (JFA's h1-a3 variation for "un": F-statistic: 53.15 on 5 and 4510 DF, p-value: < 2.2e-16). Figure 2 tracks the progress of these vocal settings across time, showing first five conversations with Chinese partners, then five with English native speakers, then ten conversations with female and male Japanese partners. We see that this selectivity is not restricted to partner alone; but that the pca coefficients also differ significantly across time. The speaker not only selects her words according to her conversational partner, but she also changes her vocal settings. Perhaps to display her relationship and attitudes towards the interlocutor.

4. DISCUSSION

This paper does not attempt to explain *why* these parameters vary as they do; that is left for future work. Instead it suffices to show that they *do* vary and that they vary both according to the interlocutor and according to progress of time through the conversations. The prosodic features have been shown to vary and in a consistent way as conversational partners become more familiar with each other across the series of recordings. We might posit, for example, that for Figure 2, a higher value indicates a more relaxed, less guarded, mode of speaking.

5. CONCLUSION

This paper has presented acoustic data showing that the four prosodic parameters, voice pitch, vocal energy, speech timing, and voice quality, vary consistently according to non-linguistic factors.

Future work will include an explanation of the mechanisms of these changes and will attempt to account for the direction and amount of change in each parameter.

6. REFERENCES

[1] Fant, G. 1960. *Acoustic Theory of Speech Production*. The Hague: Mouton.

[2] Titze I, Talkin D. 1979. "A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation". JASA 66: 60–74.

[3] Ni Chasaide, A., and Gobl, C., 1997. "Voice source variation". In W. J. Hardcastle and J. Laver (Eds.), *The Handbook of Phonetic Sciences*, pp. 428-461. Oxford: Blackwells.

[4] Campbell, N., and Mokhtari, P., 2003. "Voice Quality is the 4th Prosodic Parameter". *Proc. 15th ICPhS* Barcelona, 203–206.

[5] Hirst, D. J., 1977. *Intonative features; a Syntactic Approach to English Intonation*, The Hague, Mouton Publishers.

[6] Hanson, H. M., 1995. "Glottal characteristics of female speakers". Ph.D. dissertation, Harvard University.

[7] Käre Sjölander, 2006. The Snack Sound Toolkit from http://www.speech.kth.se/snack/

[8] Pearson, K., 1901. "On Lines and Planes of Closest Fit to Systems of Points in Space". Philosophical Magazine 2 (6): 559-572.

[9] R Development Core Team, 2004. "R: A language and environment for statistical computing". R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, http://www.R-project.org