Computational Linguistics and Chinese Language Processing
Vol. 12, No. 1, March 2007, pp. 1-16

Differences in the Speaking Styles of a Japanese Male According to Interlocutor; Showing the Effects of Affect in Conversational Speech

Nick Campbell[*+]

Abstract

There has been considerable interest recently in the processing of affect in spoken interactions. This paper presents an analysis of some conversational speech corpus data showing that the four prosodic characteristics, duration, pitch, power, and voicing all vary significantly according to both interlocutor differences and differences in familiarity over a fixed period of time with the same interlocutor.

Keywords: Conversational Speech Corpus, Expression of Affect, Prosodic Characteristics, Voice Quality Analysis.

1. Introduction

Human spoken interactions convey a variety of different types of information. In addition to the linguistic content of speech, there are also paralinguistic and extralinguistic elements that convey discourse-level and interpersonal levels of information related to the speaker, to the speaker's relationship(s) with the listener, and to the intended and actual progress of the discourse [Lindblom 1990; Stenström 1994; Hirschberg 1992, 1995].

Affect is conveyed in speech communication in a multitude of ways [Cahn 1989], including facial expression, gesture, body posture, speaking-style, tone-of-voice, lexical choice, syntactic construction, etc. It is perhaps impossible for a human to speak without revealing information about his or her affective states [Campbell 2005].

This paper examines how such affective information might be carried in the voice, particularly in the prosody of the speech, and shows from an examination of some corpus data

[*] National Institute of Information and Communications Technology
[+] Spoken Language Communication Research Laboratory, Advanced Telecommunications Research Institute International, Keihanna Science City, Kyoto 619-0288, Japan.
E-mail: nick@nict.go.jp; nick@atr.jp

[Received September 30, 2006; Revised October 30, 2006; Accepted November 1, 2006]

that evidence can be found for changes in affective state according to the nature of the interocutor and the history of their discoursal relationship.

2. The JST/CREST ESP Corpus

Speech research nowadays is predominantly corpus-based. One learns about the characteristics of speech and the expressivity of speech utterances from the analysis of a very large number of samples collected under a variety of speaking conditions [Campbell *et al*. 2006; Cowie *et al*. 2005]. For a period of five years, in order to aid the development of a technology capable of Expressive Speech Processing (ESP), the Japan Science and Technology Agency funded the collection of a large corpus of expressive speech that was coordinated by ATR in Kyoto, Japan.

As part of this corpus, over a period of three months during 2002, a group of ten volunteers were employed to talk with each other over the telephone for half-an-hour each time and to record their conversations to DAT using high-quality head-mounted condenser microphones. These conversations and their manually-produced transcriptions now form subset ESP_C of the JST/CREST ESP Corpus.

Of the ten volunteers, two were native speakers of Chinese, both fluent in Japanese, one male and one female, and two were native speakers of English, both fluent in Japanese, one male and one female. The remaining six were Japanese native speakers, three men and three women, living in the Kansai area of central Japan. They did not know each other initially but became familiar over the period of their telephone conversations. To our direct knowledge they never met face-to-face during this period.

This paper focuses on the speech characteristics of one male speaker from this corpus, JMA, who spoke with six partners over the three month period. In all, their conversations include 49,377 utterances from speaker JMA, where an utterance is approximately defined as the shortest meaningful unit of speech produced under a single intonation contour. The actual boundaries were determined on a case-by-case basis by the transcribers according to a set of rules published elsewhere [Campbell 2006].

The paper examines the acoustic characteristics of these utterances according to differences in interlocutor and stage of the interaction, showing that speaking style and voice phonation characteristics vary according to the interlocutor, in accordance with changes in familiarity and other speaker-listener relationships.

3. Materials for the Study

In previous work [Campbell 2005, 2006] affect-bearing utterances were distinguished from those that serve primarily to portray propositional or 'linguistic' content. The former, often called 'grunts' or 'affect bursts' are not usually found registered as words in a language dictionary, but are found very frequently in colloquial speech. For this study, a subset of 100 of those that occurred more than 50 times each in the conversations of one speaker (JMA) was selected, yielding 11,750 short conversational utterances for subsequent acoustic analysis. These were taken from five conversations each with each of the Chinese and English native-speakers, and from ten conversations with the Japanese native-speaker partners. Table 1 shows the number of utterances produced with each interlocutor. Table 2 lists the romanised orthographic transcriptions and counts of some of the more common examples.

Table 1. Utterance counts for the series of conversations with each interlocutor. The initial letters J,C,E in the interlocutor identifiers stand for Japanese, Chinese, and English respectively, the middle letters F and M stand for female and male respectively, and the third letter is an identifier.

CFA	CMA	EFA	EMA	JFA	JMB
1832	1632	1490	1773	2957	2066

Table 2. Counts of some of the more common utterance types that were studied for this paper, which occurred at least 50 times in the speech of JMA; the ten most common more than 1000 times. A sharp intake of breath was transcribed as '@S', a sniff as 'zu', and a laugh as '@W'. The letter 'n' indicates a syllabic nasal (umm). A minus sign represents moraic lengthening, which is known to be distinctive in Japanese, possibly triggering the percept of a different word. A dot represents a morphological boundary.

a 296	a- 368	a– 693	a— 608	a.a- 390	a-.hai 386
a.hai 577	a-.n 368	ano 337	ano- 494	a!! 927	demo 272
e- 665	e– 254	ee 2679	fun 642	fu-n 625	fu–n 273
ha.ai 978	hai 7295	ha-i 1657	hai.hai.hai 378	n(umm) 265	n- 456
n– 410	nanka 273	ne- 367	nee 284	@S 3382	sou 810
su- 429	su– 296	un 3717	u-n 2401	u–n 1243	u—n 333
un.un 351	@W 3041	zu- 1348	zu– 467		

The speech files corresponding to these utterances were analysed for their acoustic characteristics and a table of statistics for each utterance was produced. Specifically, the duration, pitch, power, and spectral characteristics of each utterance were recorded.

Duration was expressed both as absolute (log) duration of the measured utterance and as 'speaking rate' by dividing the absolute duration of the utterance by the number of phonemes in its transcription. This is a crude measure which does not take into consideration the inherent differences in different phone durations, but which serves to provide a simple approximation of speaking rate which will suffice for the present analysis.

Pitch (or more precisely, a measure of the fundamental frequency of the voice) was extracted using the ESPS 'get_f0' method that is incorporated in the 'Snack' signal processing library. The maximum and minimum pitch values for each file were recorded and stored along with an estimate of the range and average values for each utterance. The pitch contour was characterised by noting the average values measured over each third of the utterance, and stored these along with the percentage position of the pitch peak and the lowest pitch value.

Power values (*i.e,* measures of rms waveform amplitude) were calculated similarly, using the Snack command "power", and stored as maximum, minimum, average, and range for each utterance.

Spectral characteristics were calculated using the Snack command "dBPower" with options "-fftlen 128 -windowlength 128 -analysistype LPC -lpcorder 20". This produced an LPC-based 64-point vector representing the long-term average spectrum for the entire utterance (average length 0.54 seconds) from which values from points 2, 3, 4, 5, 7, and 9 were selected to represent the average power up to 1.5 kHz, points 12,15,19,23,28 to represent the average power between 1.5kHz and 4kHz, and points 34, 41, 49, 56, and 63 were selected to represent the average power between 4kHz and 8kHz. The average spectral energy measured in each of these three frequency bands was stored as a 3-valued vector for subsequent 'spectral' analysis. Since our main objective here is to examine spectral tilt, as evidence of differential phonation styles, the relative differences between the three bands (mid, high, and low on a mel-scale) were determined to suffice as a measure.

4. Analysis of the Data

After confirming independence of the variables under examination, a weak but insignificnt correlation of $r = 0.32$ was found between variations in pitch range and power range, and one of $r = 0.37$ between the averaged values of pitch and power across the 11,750 short utterances selected from the 49,377 utterances in the conversational corpus. There was a similar weak correlation between the measures of duration and power ($r = 0.34$) but none

between duration and pitch ($r = 0.19$). The correlation between spectral energy (power in the lowest band) and raw signal amplitude (signal power) was 0.08. One can thus be satisfied that the measures are sufficiently independent to carry meaningful information in their differences.

There was a clear corrrelation of $r = 0.81$ observed between energy in the first spectral band (frequencies up to 1500 Hz) and in the second (frequencies between 1.5kHz and 4kHZ), but no such correlation between frequencies in the second and third bands (*i.e.,* between frequencies below and above 4kHz) which showed a correlation of $r = 0.2$. It is the difference between these latter two bands that is of interest here, since the lack of energy in the upper frequency bands is an indicator of a less tense, more breathy, speaking style which has been shown in previous studies ([Gauffin and Sundberg 1989; Sluijter and van Heuven 1994; Campbell and Mokhtari 2003]) to correlate with intimacy and a more careful manner of speaking.

4.1 Fundamental Frequency and Power

Figure 1 shows the values of f0 measured from the speech data of the male speaker JMA plotted separately for each interlocutor. The left plot shows average f0, the middle plot maximuim f0, and the right-hand plot minimum values of f0 measured in the conversations with each interlocutor respectively. The box-plots show median and interquartile values, with whiskers extending to 1.5 times the interquartile range. The boxes are drawn with widths proportional to the square-roots of the number of observations in the groups. A notch is drawn in each side of the boxes. If the notches of two plots do not overlap this is 'strong evidence' that the two medians differ at the 5% level of confidence.

Figure 1 shows that there is more variation in the voice fundamental frequency of speaker JMA when talking to the non-native partners, while the average values of f0 for the Japanese partners JFA and JMB are higher and less dispersed. The maximum f0 is highest when speaking with the English female, and lowest when talking with the Chinese male partner. When speaking with the Japanese native speakers, the maximum f0 shows the same median values as when talking with the English female partner, but there is overall more variety in f0 when speaking with the non-native parters.

Figure 2 plots the average, maximum and minimum power values for conversations with each of the six interlocutors. It shows that more energy is used when speaking with the Japanese partners, and more variation when speaking with the non-native interlocutors. Interestingly, the minimum power appears to be higher among conversations with the Chinese male partner than among conversations with the other interlocutors, but significantly lower for conversations with the Japanese female partner.

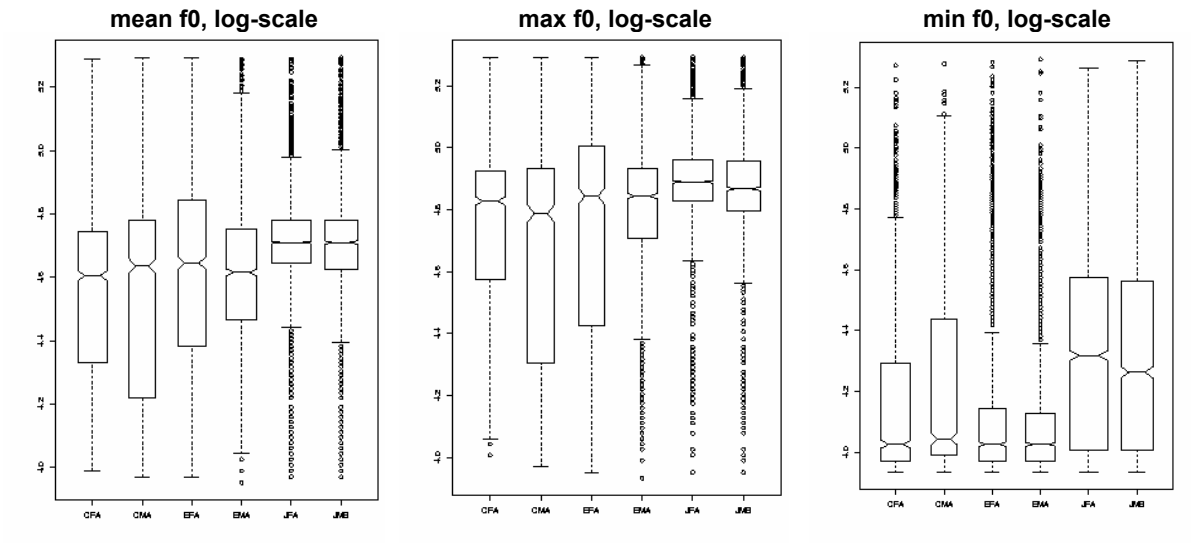

Figure 1. Plots of mean, maximum and minimum f0 values observed in the data of each of the interlocutors. The box-plots show median and interquartile values, with whiskers extending to 1.5 times the interquartile range. All F0 measurements are converted to their log values for ease of comparison.

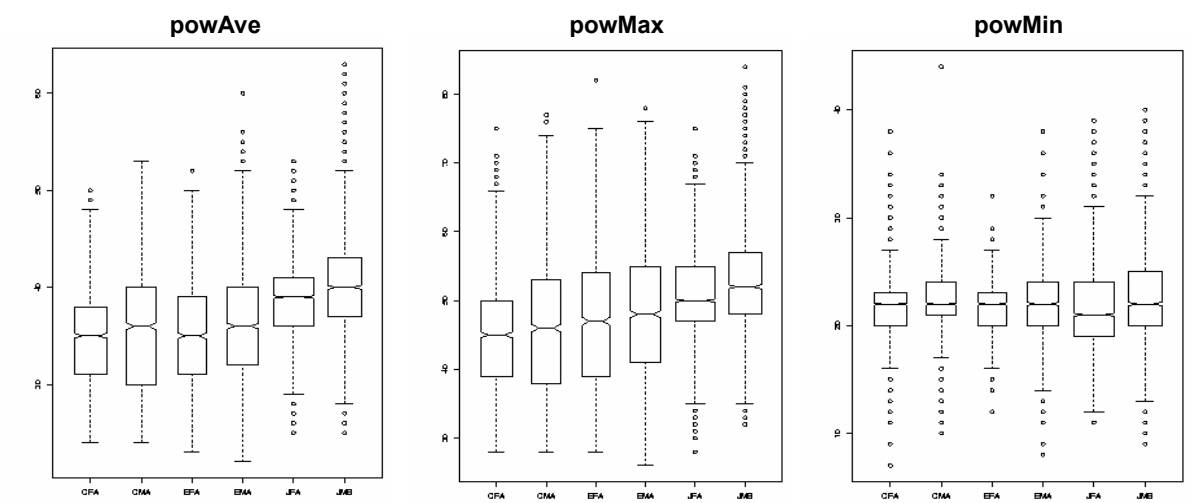

Figure 2. Plots of mean, maximum and minimum rms amplitude (speech signal power) values for each of the interlocutors.

Figure 3 plots f0 range for comparison with power range across the same set of partners. It shows a slightly higher range of f0 activity when this subject is talking with the English female than with the Chinese male partner. Both plots are log-scaled and show an average of

55Hz (*exp*(4)) median pitch range with a 30dB average power range when conversing with these different interlocutors. Power is noticeably higher when talking with Japanese native-speaker interlocutors, and lowest when talking with the female Chinese native-speaker. Needless to say, microphone distances (head-mounted) and record-level settings remained unchanged across all recordings.

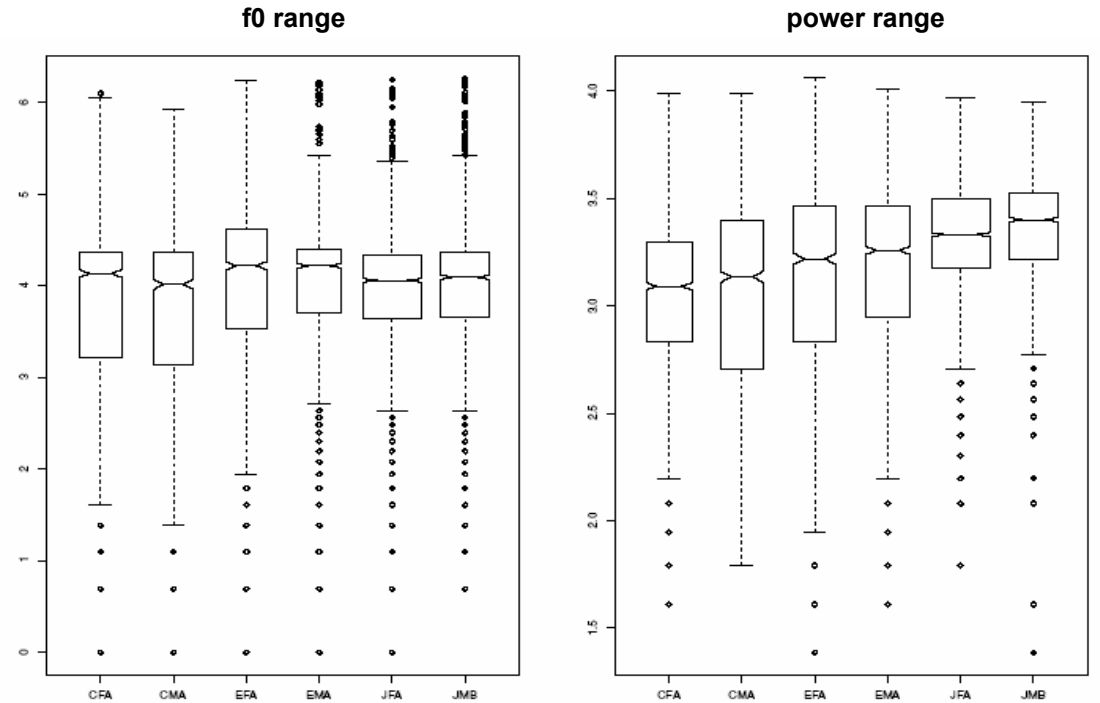

Figure 3. Ranges of fundamental frequency and power measurements for the affective utterances, plotted by interlocutor. As above, C, E, J stand for Chinese, English, Japanese respectively, and F, M represent female and male interlocutors. Both f0 and power are plotted as log values.

4.2 Duration and Speaking Rate

Figure 4 shows details of 'speaking rate' changes across the series of conversations with the two Japanese partners. This measure was calculated for each utterance by dividing the observed duration of its speech waveform (measured in milliseconds) by the number of characters in its transcription (see Table 2 for examples) and is therefore only an approximation of the true speaking rate, but it serves as a basis for comparison and provides a simple form of normalisation for the inherent differences in utterance type. Speaker JMA took part in nine conversations with female JFA, and eleven with male JMB. We note an average of 174.2 milliseconds per phone for interlocutor JFA, and an average of 169.65 for

interlocutor JMB. The speaking rate with the male partner appears to slow down throughout the series of conversations, while after reaching a peak in conversation J04 with JFA it appears to revert to a higher rate with the progression of time. Median values for JFA are 143, 162, 180, 183, 170, 166, 157, 171, and 158, while those for JMB are 144.5, 150.5, 183.0, 171.0, 149.5, 165.5, 170.5, 181.0, 167.5, 190.0, and 182.0.

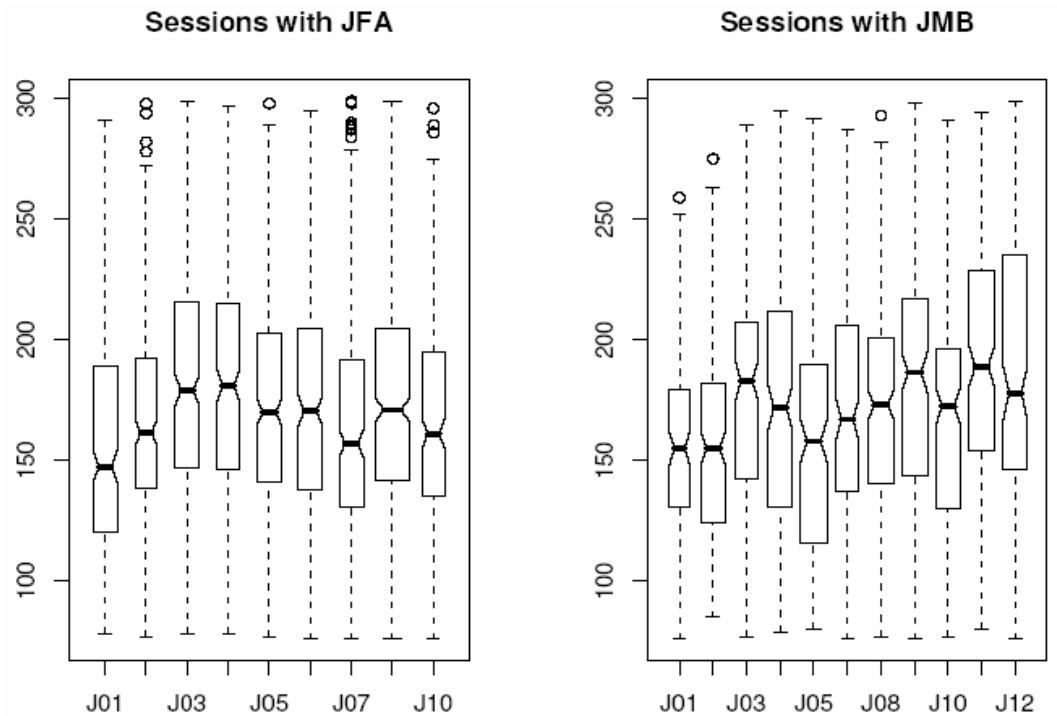

Figure 4. Speaking rate changes over weekly sessions

These values may seem unexpectedly long to an observant reader familiar with segmental durations, but it should be noted that they are sounds in afective grunts, not phones in lexical words. For example, the quantile durations (in seconds) for the word "hai" (="yes"/"I'm listening"/"I agree") are as follows: minimum: 0.152, 25th percentile: 0.382, median 0.43, 75th percentile: 0.49, and max: 1.59 seconds (n=7295). Those for the word "ee" (n=2679) are 0.268, 0.344, 0.42, 0.479, and 0.539. The durations observed for laughs in this context (n=3041) ranged from 119 milliseconds to four seconds, with a median duration of 0.9 seconds. Note that many of these utterances bear lengthening diacritics (e.g., 'a', 'a-', 'a–', 'a—', etc) and the transcribers who were all native speakers of Japanese were instructed to use one minus-sign to mark each mora-worth of lengthening perceived on the segment. It is customary to use such lengthening marks in standard Japanese Kana orthography, and mora durations are typically strictly observed in Japanese, where a moraic difference in timing

sructure can (unlike English) cue a different lexical item [1].

4.3 Spectral Slope

Spectral energy provides a simple cue to phonation style; the less energy in the higher part of the spectrum, the breathier the voice, and vice-versa. In pressed voice, the glottis closes faster as a proportion of the fundamental period, and the rapid closing is a sign of increased vocal effort and/or laryngeal muscular tension [Klasmeyer and Sendlmeier 1997; Fant 1993; Johnstone and Scherer 2000]. In conversational speech considerable use is made of voice phonation settings, especially for the display of affect [Campbell 2005; Campbell and Mokhtari 2003].

Figure 5 shows three measures of averaged spectral energy for the 11,750 affective utterances under examination. The low-frequency part of the spectrum is shown measured in decibels, as is customary in plots of spectral sections, but the middle and right-hand plots show differences between the low-frequency energy and the higher bands. Differences are also measured in decibels, and here 'low-band minus mid-band energy' is plotted in the centre plot, and 'mid-band minus high-band energy' is plotted in the right-hand plot. By plotting the differences rather than the absolute values, it is easier to visualise the spectral slope differences across these utterances.

The figure, averaged over all conversations, shows higher low-band energy for the Japanese female and the two Chinese interlocutors, with increased spectral slope for the Japanese female in the mid-band, and steeper spectral slope for the Japanese female and the two Chinese interlocutors at the top-end of the spectrum. The spectrum is therefore flatter overall for the English native-speaking partners and for the Japanese male partner. A flatter spectrum has been shown to reflect more tension in the voice.

Quantiles for the three spectral bands (measured over all data for speaker JMA) are given in Table 3, which shows median values to be -42, 14, and 9 decibels respectively. The difference of 14 decibels indicates that the average value for energy measured in the frequency range between 1.5kHz and 4kHz is -56 decibels, while the energy between 4kHz and 8kHz is typically at the -65 decibel level.

---

[1] For example, in Japanese, 'ie' means "house" while 'iie' with a longer first vowel means "no". Similarly, 'ka' is an interrogative particle, while 'ka-' with a lengthened vowel means "car". Such length-based lexical distinctions are common.

Table 3. Quantiles of the energy measured in three spectral bands. The top row shows absolute energy but the bottom two rows show energy differences measured in decibels.

	0%	25%	50%	75%	100%
low-band	-72	-48	-46	-44	-21
mid-band	-11	12	14	16	38
high-band	-18	7	9	11	37

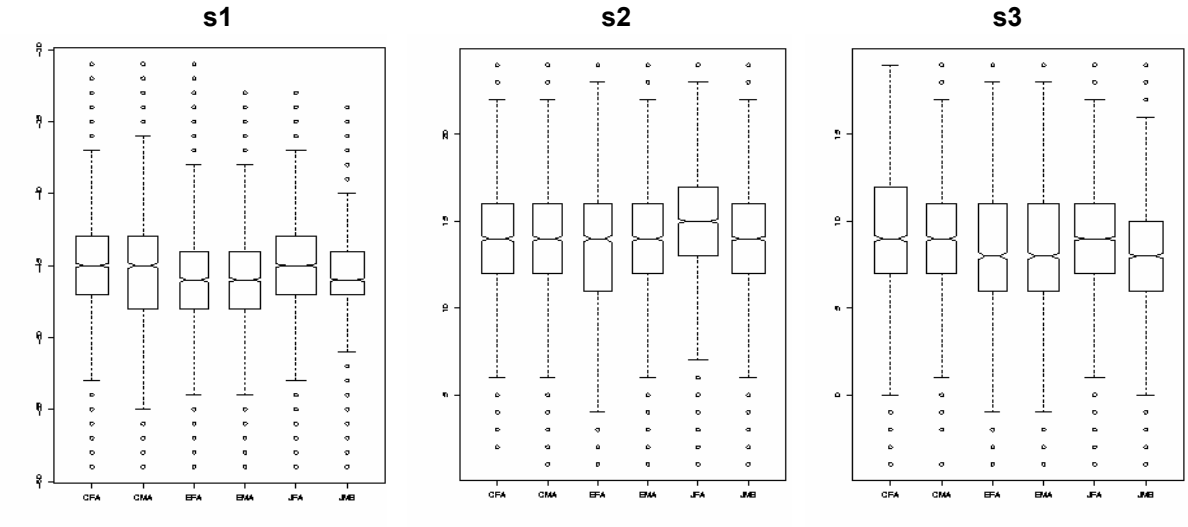

Figure 5. Three measures of spectral energy provide an indication of spectral slope. The left-hand plot shows average energy measured between 0-1.5kHz, the middle plot the difference between that and average energy measured between 1.5kHz and 4kHz, and the right-hand plot shows the difference between the averaged mid-band energy and the averaged energy between 4kHz and 8kHz at the top end of the spectrum. Measures are plotted separately by interlocutor.

Figure 6 shows differences in these values over time. The upper three plots show low-band, mid-band, and high-band energy measures for conversations with Japanese male partner JMB. The lower part of the figure shows only the high-band energy differences (spectral slope measures) for the four non-native partners.

In each case there is a general trend towards decrease in steepness of the spectral slope with time. From the top plots (of the series of conversations with partner JMB), the second, third and penultimate conversation exhibited high low-frequency spectral energy (from the left-hand plot), steep falloff in mid-band energy (from the central plot), and, at least for the

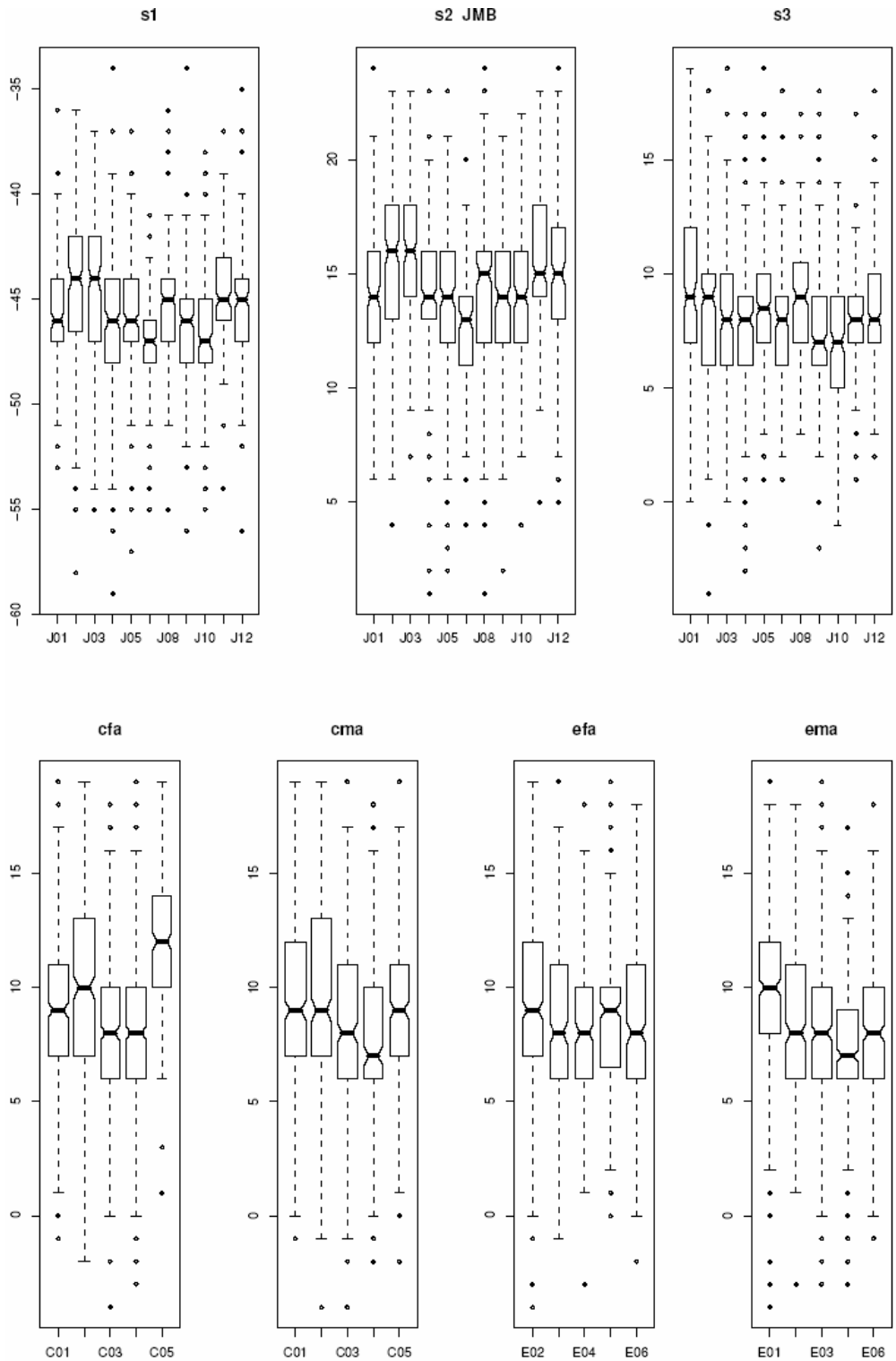

Figure 6. Spectral slope differences across time by interlocutor. The top part shows all three spectral bands for partner JMB, and the bottom part shows the difference between mid-band and high-band energy for each of the non-native partners.

second and third conversations, considerable variability in the high-frequency dropoff. This would be consistent with a higher degree of tension and varying politeness in the speech of the initial and penultimate conversations. For conversations in the interim period, from J04 to J10, a gradual decrease of steepness is found in the high-end spectral tilt that would be consistent with an increase in familiarity as reflected by more frankness and less polite softening of the voice. Then for the final conversations, as the recordings (and the three-month relationship) come to an end, there is an increase again, as would be consistent with a rise in formality of the conversational speech between the partners.

In the lower plots, with the non-native speaker partners, there is a similar steeepness in high-frequency dropoff (consistent with increased politeness in the voice and speaking style) in the initial and final recordings, and a gradual relaxation of spectral tilt in conversations of the interim period. Steeper spectral slope is found in the conversations with the female partners, with the Chinese female being highest and the English male being lowest in this respect.

5. Discussion

In this analysis of the prosodic characteristics of the conversational speech of one Japanese male over a period of three months, considerable variation was found in all of the parameters measured. By factoring the analysis according to differences in interlocutor as well as by differences in time, or sequence of the conversations, we were able to show that the changes are not a result of time-related changes, such as tiredness or ill-health, but that they correlate more with differences in interlocutor and with development of the individual relationships.

It is probable that not all interlocutors were related to in an equal way. One can imagine more sharing of common interests between native speakers of the same language, and different forms of bonding in the relationships that developed between the male and the female partners respectively. Similarly, the culturally closer, Asian but foreign, Chinese partners and the possibly exotic, and maybe more foreign, English speakers would have brought different contributions and cultural assumptions to the conversations. Their necessary lack of fluency in the use of Japanese, particularly over the telephone where the visual support for communication is impaired, would have introduced idiosyncracies into the style of the different conversations.

Without a complementary analysis of the texts of the conversations, one can only draw speculative conclusions to explain the differences in the prosodic characteristics, but from the spectra of speech with partner JMB in Figure 6 it can be assumed that the initial relatively low spectral energy and high spectral tilt of conversation J01 represent the ‘baseline’ settings for speaker JMA who had no expectations at that time about his partners. One might then

speculate that the apparent increase in politeness (as indicated by a more breathy speaking style) in conversations II and III could be due to having to maintain a conversation for a long 30-minutes over the telephone with a partner who is still relatively unknown to the speaker, and that the decrease thereafter occurred as they found more interests in common to talk about. From a brief examination of the transcriptions, they certainly appear to have become friends over the three month period. If so, then perhaps one can also speculate that the increase of breathiness in their speech towards the end is indeed due to the approaching termination of their telephone relationship.

6. Conclusion

In light of the recent considerable interest in the processing of affect in spoken interactions, an analysis was performed of some corpus data of conversational speech, showing that the four prosodic characteristics, duration, pitch, power, and voicing all vary significantly according to interlocutor differences and to differences in familiarity and politeness over a fixed period of time with the same interlocutor.

The results showed significant differences in the prosodic characteristics of speech with others sharing the same native language as compared with those of non-native speakers of Japanese. The results also showed that speaking rate, pitch range, and spectral tilt varied significantly according to partner and position of the conversation in the three-month series. Because different settings were used with different partners at the same time, the possibility can be discounted that these differences were due to unrelated external considerations such as variation in the health of the speaker.

The findings reported earlier for similar changes in phonation settings for a female speaker from a separate section of the corpus (see [Campbell 2005; Campbell and Mokhtari 2003]) under more varied conversational settings have been replicated here with data from a different speaker in a more controlled recording environment.

It is perhaps still too early to make use of these findings in speech technology, and considerable further work is required before strong claims can be made about the causes and relationships, but it is of interest that these differences exist at all. Listeners certainly make use of small but consistent speaking-style and phonation-setting changes to make inferences about the affective states of the speaker. Perhaps these variations will provide the foundation for both speech synthesis and speech recognition modules that begin to incorporate affect as one of the strands of meaning in speech. Such technology would be of great use in providing a softer interface between machines and humans in society.

References

Cahn, J., "The generation of affect in synthesised speech," *Journal of the American Voice I/O Society,* Vol c8, 1989, pp. 251-256.

Campbell, N., "Getting to the heart of the matter; speech as expression of affect rather than just text or language," *Language Resources & Evaluation,* 39(1), Springer, 2005, pp. 109-118.

Campbell, N., "Conversational Speech Synthesis and the Need for Some Laughter," *IEEE Transactions on Audio, Speech, and Language Processing,* 14(4), July 2006.

Campbell, N., "A Language-Resources Approach to Emotion: Corpora for the Analysis of Expressive Speech," In *Proc International Conference on Language Resources and Evaluation*, LREC 2006.

Campbell, N., L. Devillers, E. Douglas-Cowie, V. Auberge, A. Batliner, and J. Tao, "Resources for the Processing of Affect in Interactions," Panel session, In *Proc LREC'06*, Genoa, Italy, 2006, pp. xxiv-xxvii.

Campbell, N., and P. Mokhtari, "Voice Quality; the 4th prosodic parameter," In *Proc 15th ICPhS*, Barcelona, Spain, 2003.

Cowie, R., E. Douglas-Cowie, and C. Cox, "Beyond emotion archetypes; Databases for emotion modelling using neural networks," In *Neural Networks 18*, 2005, pp. 371-388.

Fant, G., "Some problems in voice source analysis," *Speech Communication,* 13(1), 1993, pp. 7-22.

Gauffin, J., and J. Sundberg, ''Spectral correlates of glottal voice source waveform characteristics," *Journal of Speech and Hearing Research,* 32, 1989, pp. 556-565.

Hirschberg, J., ''Using discourse content to guide pitch accent decisions in synthetic speech," In G. Bailly and C. Benoit, ed, *Talking Machines*, North-Holland, 1992, pp. 367–376.

Hirschberg, J., ''Acoustic and prosodic cues to speaking style in spontaneous and read speech," In *Symposium on speaking styles, Proc ICPhS*, Stockholm, Sweden. 1995.

Johnstone, T. and K. R. Scherer, "Vocal Communication of Emotion," in: M. Lewis & J. Haviland (Eds.) Handbook of Emotion (2nd ed.). New York: Guildford. 2000.

Klasmeyer, G., and W. F. Sendlmeier, "The classification of different phonation types in emotional and neutral speech," *Forensic Linguistics,* 4(1), 1997, pp. 104-124.

Lindblom, B. E. F., ''Explaining phonetic variation: A sketch of the H&H theory," In *Speech Producstion and Speech Modelling, NATO-ASI Series D: Behavioural and Social Sciences,* edited by H. J. Hardcastle and A. Marchal (Kluwer, Dordrecht), Vol 55, 1990.

Schroeder, M., "Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions," In *Proc. Workshop on Affective Dialogue Systems: Lecture Notes in Computer Science*, Kloster Irsee, Germany, 2004, pp. 209-220.

Sluijter, A. M. C., and V. J. van Heuven, "Spectral tilt as a clue for linguistic stress," presented at 127th ASA, Cambridge, MA. 1994.

Stenström, A., *An Introduction to Spoken Interaction*. Longman, London. 1994.

Website Resources

Snack: a Tcl/Tk library and toolkit for speech signal processing. http://www.speech.kth.se

The Japan Science & Technology Agency *Core Research for Evolutional Science & Technology*, 2000-2005.

The JST/CREST Expressive Speech Processing Project homepage can be found at http://feast.atr.jp/