authentic Japanese syllables. In this case it would seem that the application of the duration rules for Swedish quantity could have yielded a rather good rendering of the Japanese contrast. These results indicate that in the case of the realization of Japanese quantity, the transfer of at least some of the aspects the Swedish quantity contrast pattern is part of the Swedes' strategy in learning Japanese quantity. The durational aspects of the English tense-lax contrast present a somewhat less clear picture of the transfer phenomenon. It looks like the Swedish natives are attempting to render the contrast but could be unsuccessful because of their tendency to continue to apply the L1 pattern in their L2 use.

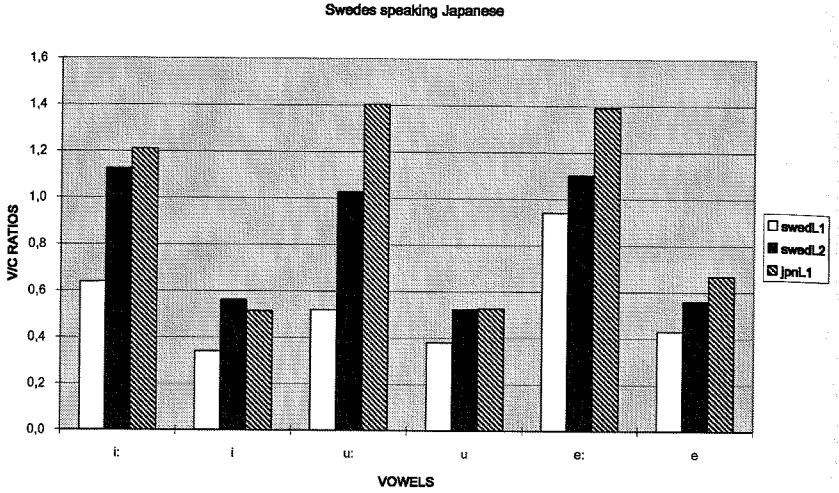

Figure 2 shows the calculated V/C duration ratios for the short-long-vowels in Japanese averaged over both isolated words and words which occurred in a sentence.

Further work on this material can give us a clearer idea of what residue from the L1 there might be in the phonetic realization of an L2 contrast.

References

Elert, C-C., 1964. *Phonologic Studies of Quantity in Swedish.* Uppsala: Monografier utgivna av Stockholms kommunalförvalting 27.

Flege, J. & W. Eefting, 1986. The production and perception of English stops by Spanish speakers of English. *Journal of Phonetics 15*, 67-83.

McAllister, R., J.L. Flege & T. Piske, 2003. The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian. *Journal of Phonetics 30*, 229-258.

Schaeffler, F., 2005. Phonological Quantity in Swedish Dialects. *PHONUM 10.*

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 97–100

Cross-speaker Variations in Producing Attitudinally Varied Utterances in Japanese

Yasuko Nagano-Madsen[1] and Takako Ayusawa[2]
[1]Department of Oriental and African Languages, Göteborg University
yasuko.madsen@japan.gu.se
[2]Department of Japan Studies, Akita International University
ayusawa@aiu.ac.jp

Abstract

Several acoustic phonetic parameters were analysed for six professional speakers of Japanese who produced attitudinally-varied utterances. The results showed both agreement and discrepancies among the speakers, implying that pragmatic information can be expressed in at least a few alternative ways in Japanese and that this line of research needs more attention.

1 Introduction

It is well known that pragmatic information can be combined in a set of tunes (or pitch-accents in more recent terminology) in a language like English which has been traditionally called an intonational language. How such pragmatic information is conveyed in a tone or pitch-accent language in which pitch shape is lexically determined is much less clear. For Japanese, Maekawa & Kitagawa (2002) conducted pioneering research on the production and perception of paralinguistic phenomena. We have earlier reported the F0 shape characteristics to show how speakers choose pitch shapes and phrasing to convey pragmatic meanings in Japanese (Nagano-Madsen & Ayusawa, 2005). In this paper, we will report other phonetic cues used by the same speakers. The attitudes tested are NEU(tral), DIS(appointment), SUS(picious), JOY, and Q(uestion). Three phonologically balanced short utterances were produced as a reply by six speakers – three male and three female speakers. For details on data, speakers, and procedure, see Nagano-Madsen & Ayusawa (2005).

2 F0 characteristics

2.1 Pitch range

In order to make the cross-speaker comparison more meaningful, F0 features are calculated on a semitone scale rather than in absolute Hz values. The average pitch ranges for the female and male speakers were 13.9 and 14.3 semitones respectively. Table 1 shows the average pitch range in semitones for the six speakers for the five attitude types, which shows that the overall average pitch range increases in ascending order, DIS<NEU<SUS<Q<JOY. Speakers were uniform in using their narrowest pitch range, on average 10.9 semitones, in expressing attitude DIS. The fact that attitude DIS had the narrowest pitch range agrees with the findings reported in Maekawa & Kitagawa (2002), though the exact magnitude of range cannot be compared with their data. The widest pitch range was used for JOY, with an average of 16.3 semitones. Considerable cross-speaker variation is found in the use of overall pitch range indicating that the overall pitch range alone cannot be regarded as a reliable acoustic phonetic cue for attitude types. The male speakers manifest pitch range for NEU and JOY more closely

than female speakers. It would be interesting to know if the three male speakers instead use other phonetic cues more actively than female speakers.

Table 1. Cross-speaker variation in F0 range for attitude (F0 maxima minus (final) F0 minima in semitones).

	Female speakers			Male speakers			All the speakers
Attitude/ speaker	U	V	W	X	Y	Z	
Q	11.4	14.7	14.5	16.0	15.8	17.2	15.0 (2.68)
SUS	12.0	19.2	17.7	12.0	13.4	14.7	14.9 (3.27)
JOY	15.5	16.3	20.1	13.2	17.1	15.5	16.3 (3.73)
DIS	10.8	10.8	10.4	11.3	9.5	12.3	10.9 (1.66)
NEU	12.5	13.1	12.4	14.5	15.6	16.4	14.1 (1.94)

2.2 Pitch range for initial rise and final fall

The pitch range was calculated for the initial F0 rise and final F0 fall (cf. Figure 1 below). A typical manifestation of the pitch range of initial F0 rise is in ascending order DIS<NEU<Q<SUS<JOY. The cross-speaker pitch range variation for the initial F0 rise is far more consistent than that of the final fall in relation to attitude type. Note that there is a considerable cross-speaker variation in the manifestation of pitch range for the F0 fall for attitudes SUS and JOY, but not for the initial F0 rise.

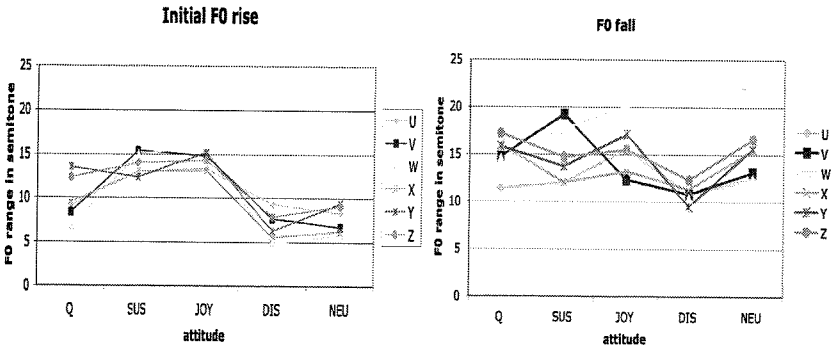

Figure 1. Pitch range in semitones for the initial F0 rise (left) and for the F0 fall (right).

2.3 Pitch range for final rise (Q and SUS)

Question utterances in Japanese are typically accompanied by a terminal rising contour. In the present data, even SUS utterances had a regular terminal rise. However, final F0 rise for Q and SUS were consistently differentiated in the magnitude (cf. Figure 2). The average F0 rise for Q was 9.1 semitones (SD=3.04) while that for SUS was 12.9 semitones (SD=4.72). The magnitude clusters around 2-4 semitones for most speakers, but speakers U and W had more extreme values.

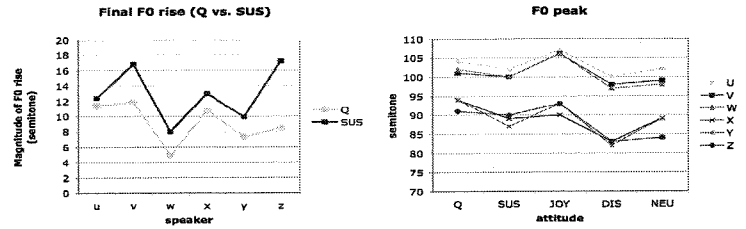

Figure 2 (left). Magnitude of utterance final F0 rise in semitone for Q and SUS.
Figure 3 (right). F0 peak values for the six speakers. U, V, W are female speakers.

2.4 F0 peak value

Figure 3 shows the F0 peak values for different attitude types. Four out of six speakers had the same order in the F0 peak value, which in ascending order is DIS<NEU<SUS<Q<JOY. Note that this order is similar to that of the pitch range of initial F0 rise except that of SUS and Q. Two speakers, both male, had their highest F0 peak for Q rather than JOY.

2.5 F0 peak delay

The relevance of the F0 peak delay, i.e. the F0 peak is not on the expected syllable to which the phonological accent is affiliated to, has been discussed for some time in relation to pragmatics. In the present data, the F0 peak delays were common even for NEU (cf. Figure 4). All except of one case (speaker X for NEU) had a peak delay of varying from one mora-delay to six morae-delay. All the speakers had theleast peak delay for NEU while the peak delay in relation to other attitude types varied considerably across speakers with SUS showing most agreement in delay. Since the diversity among the speakers is great, it seems that the F0 peak delay per se is not a reliable correlate for attitude types.

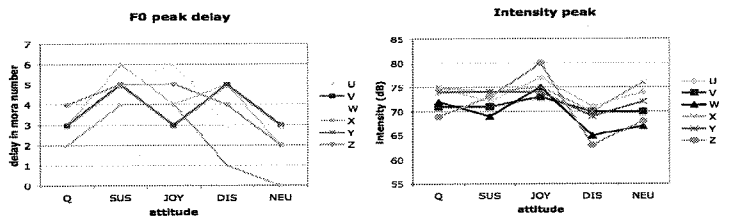

Figure 4 (left). The timing of F0 peak with the mora. When it is 0, the F0 peak is in the syllable (mora) to which the accent is phonologically affiliated and there is no delay.
Figure 5 (right). Intensity peak measurements in dB.

3 Intensity peak

There was a good cross-speaker agreement among the speakers in the intensity peak value with attitude type. The highest intensity peak value (average 75dB) was found for JOY while the highest (average 68dB) for DIS (cf. Figure 5 above). Intensity peaks in relation to the attitude types varied less across speakers. In contrast to the difference between JOY and DIS, variations in the intensity peak value for other types of attitudes is small (71-2 dB on average). However, speakers differ considerably in the magnitude of intensity peaks. Some speakers vary the intensity greatly for attitude types (speaker W and Z) while speaker U hardly varied

it. Intensity peaks and F0 peaks correlate to some extent, yet it is clear that the two parameters should be treated separately. Note that speakers W and U have very similar F0 peak values but different intensity peak values.

4 Duration (speaking rate) and pause

Average total utterance duration for the three utterances for each speaker is presented in Figure 6. Of the three utterances, the utterance /a-soo-desuka/ permits the insertion of a pause after the initial interjection /a/. When pause duration is included, it shows the same durational pattern as the other two utterances without a pause in reflecting the attitude types. Therefore, we interpreted pause as part of durational manifestation and included it in total utterance duration. The smallest cross-speaker variation was found for NEU for which all except one speaker used the shortest duration, clustering around 600-800ms. In the absolute duration value, speakers were also uniform for SUS which falls in the range between 1000 to 1200ms. The greatest cross-speaker variation was found for DIS for which the duration of the utterance varied from 800ms to 1250ms.

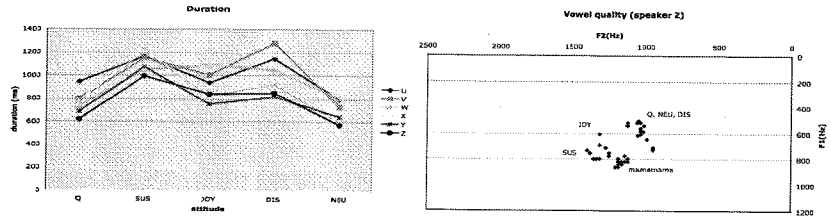

Figure 6 (left). Average utterance duration for each attitude type.
Figure 7 (right). Plotting of F1 and F2 for the vowel /a/ (speaker Z).

5 Vowel quality

Auditory impressions suggested considerable intra- and cross-speaker variation in the use of voice quality in general as well as in the specifically tested attitude types. Since the acoustic cues for voice quality are less straightforward than other acoustic cues, we only present the differences in vowel quality in this paper. Figure 7 above shows the manifestation of vowel quality by speaker Z. This speaker differentiated the vowel quality of /a/ in such a way that SUS and JOY had a more front quality than NEU, Q, and DIS. The figure also shows the formants values of /a/ in nonsense words /mamamama/ spoken neutrally by the same speaker.

6 Summary and discussion

Together with our earlier report on F0 shape and phrasing (Nagano-Madsen & Ayusawa 2005), both agreement and discrepancies were observable among the six speakers in their manifestation of attitudes. It seems that pragmatic information can be expressed in at least a few alternative ways in Japanese and that this line of research needs more attention.

References

Maekawa, K. & N. Kitagawa, 2002. How does speech transmit paralinguistic information? (in Japanese). *Cognitive Studies 9*(1), 46-66.

Nagano-Madsen, Y. & T. Ayusawa, 2005. Prosodic correlates of attitudinally-varied back channels in Japanese. *Proceedings FONETIK 2005*, Department of Linguistics, Göteborg University, 103-106.

Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics
Working Papers 52 (2006), 101–104

Emotion Recognition in Spontaneous Speech

Daniel Neiberg[1], Kjell Elenius[1], Inger Karlsson[1], and Kornel Laskowski[2]
[1]Department of Speech, Music and Hearing, KTH, Stockholm
{neiberg|kjell|inger}@speech.kth.se
[2] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA
kornel@cs.cmu.edu

Abstract

Automatic detection of emotions has been evaluated using standard Mel-frequency Cepstral Coefficients, MFCCs, and a variant, MFCC-low, that is calculated between 20 and 300 Hz in order to model pitch. Plain pitch features have been used as well. These acoustic features have all been modeled by Gaussian mixture models, GMMs, on the frame level. The method has been tested on two different corpora and languages; Swedish voice controlled telephone services and English meetings. The results indicate that using GMMs on the frame level is a feasible technique for emotion classification. The two MFCC methods have similar performance, and MFCC-low outperforms the pitch features. Combining the three classifiers significantly improves performance.

1 Introduction

Recognition of emotions in speech is a complex task that is furthermore complicated by the fact that there is no unambiguous answer to what the "correct" emotion is for a given speech sample (Scherer, 2003; Batliner et al., 2003). Emotion research can roughly be viewed as going from the analysis of acted speech (Dellaert et al., 1996) to more "real", e.g. from automated telephone services (Blouin & Maffiolo, 2005). The motivation of this latter is often to try to enhance the performance of such systems by identifying frustrated users.

A difficulty with spontaneous emotions is in their labeling, since the actual emotion of the speaker is almost impossible to know with certainty. Also, emotions occurring in spontaneous speech seem to be more difficult to recognize compared to acted speech (Batliner et al., 2003). In Oudeyer (2002), a set of 6 features selected from 200 is claimed to achieve good accuracy in a 2-person corpus of acted speech. This approach is adopted by several authors. They experiment with large numbers of features, usually at the utterance level, and then rank each feature in order to find a small golden set, optimal for the task at hand (Batliner et al., 1999).

Classification results reported on spontaneous data are sparse in the literature. In Blouin & Maffiolo (2005), the corpus consists of recordings of interactions between users and an automatic voice service. The performance is reported to flatten out when 10 out of 60 features are used in a linear discriminant analysis (LDA) cross-validation test. In Chul & Narayanan (2005), data from a commercial call centre was used. As is frequently the case, the results for various acoustic features were only slightly better than a system classifying all exemplars as neutral. Often authors use hundreds of features per utterance, meaning that most spectral properties are covered. Thus, to use spectral features, such as MFCCs, possibly with additional pitch measures, may be seen as an alternative. Delta MFCC measures on the utterance level have been used earlier, e.g. in Oudeyer (2002). However, we have chosen to model the distribution of the MFCC parameters on the frame level in order to obtain a more detailed description of the speech signal.