LACUS FORUM XXIX

Linguistics and the Real World

THE UNIVERSITY OF TOLEDO

REPRINT

ACOUSTIC REALIZATION OF PROSODIC TYPES: CONSTRUCTING AVERAGE SYLLABLES

Masahiko Komatsu
University of Alberta
Sophia University, Japan

Takayuki Arai
Sophia University, Japan

languages are classified in terms of accent (stress accent, pitch accent, and tone accent) and rhythm (stress-timed, syllable-timed, and mora-timed)[1]. Although there have been comparative studies of specific languages, they have not offered the whole picture of prosody including both accent and rhythm in terms of purely acoustic parameters, or the 'acoustic map of prosody' (Figure 1, overleaf).

Previous studies have not provided a clear acoustic definition of prosodic types. For example, concerning accent types, Eady (1982) compared the pitch of Chinese (a tone accent language) and English (a stress accent language). Because he analyzed only these two languages, pitch accent (e.g. Japanese) is not taken into consideration, and it is not clear whether his results reflect only accent types or they are also affected by difference in rhythm between the two languages. It should also be mentioned that his analysis is limited to only global characteristics of pitch. As to rhythm types, Ramus et al. (1999) proposed the variations of consonant durations and the proportion of vowel durations in speech as measures of rhythm types, based on the analysis of several languages, including English, Spanish, and Japanese. However, their analysis is not purely acoustic because they assume phonological segments, and phonological judgment intervenes in identifying consonant and vowel intervals in the speech signal. Traditionally, the acoustic correlates of rhythm types are controversial.

This paper seeks to identify acoustic measures quantitatively representing the accent and rhythm types, investigating samples from Chinese, English, Japanese, and Spanish. The paper presents two proposals on the acoustic analysis of prosodic types. The first one deals with 'what to analyze.' We propose that pitch, intensity, and Harmonics-to-Noise Ratio (hnr) should be analyzed. Such *acoustic elements* constitute the characteristics of prosodic types. Next, we propose 'how to analyze.' While some of the *global characteristics* predicted from previous studies seems effective; simple statistical calculations of acoustic elements, such as mean and standard deviation, do not necessarily characterize prosodic types. Specifically, we emphasize the importance of *local characteristics*, such as histograms of rate of change in pitch, stylized pitch line segments, instantaneous harmonics-noise plots, and average syllables. (Henceforth, global and local characteristics are called *acoustic characteristics*.) Although we have not reached a definite answer to the question of what the acoustic correlates of prosody are, we propose several prospective methods to analyze such characteristics.

Reprinted from *LACUS Forum XXIX: Linguistics and the Real World*, edited by Douglas W. Coleman, William J. Sullivan & Arle Lommel. 2003. Houston: LACUS.

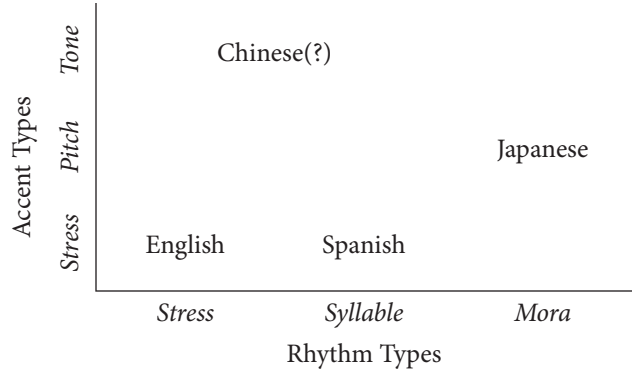

Figure 1. *Image of an 'acoustic map of prosody.' Each dimension should be related to one or more acoustic parameters. The classification shown in the figure is tentative and should be further investigated.*

1. DATA AND SIGNAL PROCESSING.

1.1. DATA. Speech data used in this paper, except the samples in Section 4.1, were taken from a multilingual speech corpus (Itahashi 2002). Of the eleven languages the corpus includes, Chinese, English, Japanese, and Spanish were used. The material used was the reading of the text 'The wind and the sun' in each language, which was approximately 30 seconds long on average. The text was read by native speakers of each language: Chinese 6 males and 7 females, English 3 males and 4 females, Japanese 6 males and 6 females, and Spanish 4 males and 5 females. All of these data were used unless otherwise noted. The data format was 16 kHz, 16-bit.

1.2. ANALYSIS. Praat Version 4.0.18 (Boersma & Weenink 2002) was used to extract pitch, intensity, and Harmonics-to-Noise Ratio (HNR) from the speech signal. The analysis was carried out frame by frame at a time step of 10 ms. (Henceforth, each analyzed section in speech is called a frame; in other words, frames were shifted by 10 ms in the analysis procedure.) Stylization of pitch by line segments was also conducted using Praat. The output of Praat was further processed for descriptive statistical calculation and graphic representation mainly by scripts in Matlab. If the frames that did not have an amplitude above 1/10 of the global maximum continued for 200 ms or more, they were regarded as pauses and eliminated from the analyses.

Because this is a pilot study, errors in pitch estimation by Praat were not modified. We consider that the errors did not have significant effects on our analyses because they were not frequent.

1.3. SYNTHESIS. We synthesized four types of speech samples from the original speech for the discussion in Section 2:

1. Sounds that retain only the pitch of the original speech. These are expected to carry information on accent and intonation.

Reprinted from *LACUS Forum XXIX: Linguistics and the Real World*, edited by Douglas W. Coleman, William J. Sullivan & Arle Lommel. 2003. Houston: LACUS.

2. Sounds that retain only intensity.
3. Sounds that retain intensity and HNR. The combination of intensity and HNR is expected to provide perceptual cues to phoneme classes.
4. Sounds that retain all three, i.e. pitch, intensity, and HNR.

All of these were created by a script of Praat. The script first calculated pitch, HNR, and intensity of the original speech. Then it created the four types of sounds based on these measures.

The sound with only pitch is simply a pulse train with a constant amplitude. Pitch was interpolated where it was not available in the original signal, e.g. for voiceless consonants. As noted above, if amplitude less than the threshold continued for 200 ms or more in the original signal, this section was regarded as a pause and suppressed to silence in the output sound. Finally, the sound was de-emphasized, i.e. tilted –6 dB per octave, to make it sound like a human voice.

The sound with only intensity was made by being driven by white noise. It was de-emphasized as well.

The sound with intensity and HNR is a mixture of a pulse train and white noise. From intensity and HNR, the amplitudes of a harmonics component and a noise component of the original signal were calculated respectively. Then, a pulse train was made based on the amplitude of the harmonics, white noise was made based on the amplitude of noise, and they were added together. The pulse train had a flat pitch. Finally, the sound was de-emphasized. The process of de-emphasizing caused a small change in HNR, but this did not make much audible difference.

The sound with pitch, intensity, and HNR is the same as the above sound except that the pulse train was made so as to keep the pitch of the original sound.

All of these sounds have a spectrum smoothly declining at –6 dB/oct. The sound with pitch, intensity, and HNR corresponds to the 'source' of a vocoder or the source-filter theory.

2. ACOUSTIC ELEMENTS. We propose that pitch is associated with accent types and that intensity and HNR are associated with rhythm types.

Accent types are expected to be characterized by pitch, following Eady (1982).

Rhythm types are, according to Ramus et al. (1999), characterized by the variations of consonant durations, the proportion of vowel durations, etc. We adopted intensity and HNR as acoustic measures indicating the degree of 'consonantal' and 'vocalic.' Intensity is closely related to sonority. Vocalic sections in speech have greater intensity, and consonantal, less intensity. HNR is the ratio of the amplitude of harmonics (i.e., periodic components) to the amplitude of noise (i.e., non-periodic components). Vocalic sections have a higher HNR, and consonantal, a lower HNR.

We synthesized the four types of sounds that include the above acoustic elements, as described in Section 1.3, for the first 10 seconds of speech of one male and one female from each language. As a trial, we used only one male and one female, and this proved the validity of the technique. It was confirmed that we can distinguish one

Reprinted from *LACUS Forum XXIX: Linguistics and the Real World*, edited by Douglas W. Coleman, William J. Sullivan & Arle Lommel. 2003. Houston: LACUS.

language from another (i.e., one prosodic type from another) by listening to these sounds. (The relation of the combination of acoustic elements and the identifiability of specific prosodic types is our future research theme.)

3. ACOUSTIC CHARACTERISTICS: GLOBAL AND LOCAL.

3.1. PITCH.

3.1.1. REPLICATION OF EADY'S EXPERIMENT (1982). Several values shown in Eady (1982) were calculated for our data: DURATION (the total duration of the speech signal), MEANFF (the mean F0 for voiced speech), SDFF (the standard deviation of F0 for voiced speech), RCFF (the average rate of change in F0 for every 10-ms interval of voiced speech), and FLUXSEC (the average number of fluctuations per second in the F0 pattern). Of these parameters, Eady's result showed a difference in MEANFF, RCFF, and FLUXSEC between Chinese and English.

Our result generally conformed to Eady's, though not perfectly, and it also gave additional insights. RCFF showed a difference between Chinese and English, as in Eady's result, and the values of Japanese and Spanish were located between Chinese and English. This suggests that RCFF is an important characteristic of accent types. FLUXSEC showed a difference between Chinese and English, but the differences between speakers' genders in Japanese and Spanish were equally large. It should also be noted that FLUXSEC seems to be susceptible to its calculation algorithm. MEANFF and SDFF did not show a clear difference across languages. DURATION showed a clear difference across languages, but it is likely to be the reflection of the lengths of the texts read rather than of prosodic types. It should be noted that RCFF and FLUXSEC are related to local shapes of pitch contours, although they are globally averaged values. The results together suggest that how pitch changes locally is important rather than its global values.

3.1.2. HISTOGRAMS OF RATE OF CHANGE IN PITCH. Because the importance of local characteristics was suggested, we scrutinized RCFF more in detail, even though Eady gave only the average values of RCFF. Figure 2 shows the distributions of RCFF. (We used 3 males and 3 females for each language to balance the number of speakers represented in each graph.) The asymmetry of positive and negative values is seen in all of the four languages, which indicates that pitch falling is more frequent than pitch rising in any language. Another important observation is that this asymmetry is larger in Chinese and Japanese than in English and Spanish. This may be relevant to the fact that pitch bears the function of distinguishing words in Chinese and Japanese.

3.1.3. DISTRIBUTION OF STYLIZED LINE SEGMENTS OF PITCH. Further, local shapes of pitch change were investigated. Original pitch contours were stylized by line segments (resolution: 2 semitones), as illustrated in Figure 3. Figure 4 shows the distribution of stylized line segments (3 males and 3 females for each language). Each dot represents a line segment. Dots with larger values of time (x-axis) indicate longer line segments; that is, pitch changes slowly or does not change for a long period of time. Dots with

Reprinted from *LACUS Forum XXIX: Linguistics and the Real World*, edited by Douglas W. Coleman, William J. Sullivan & Arle Lommel. 2003. Houston: LACUS.

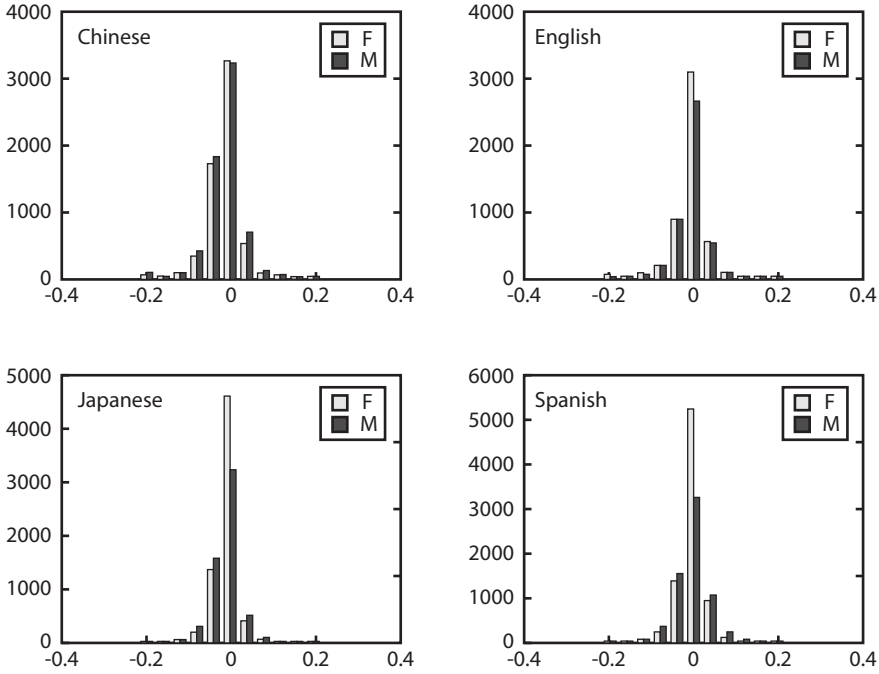

Figure 2. *Histograms of* RCFF. X*-axis: Pitch change per 10 ms (log₂[Hz]);* Y*-axis: Number of frames.*

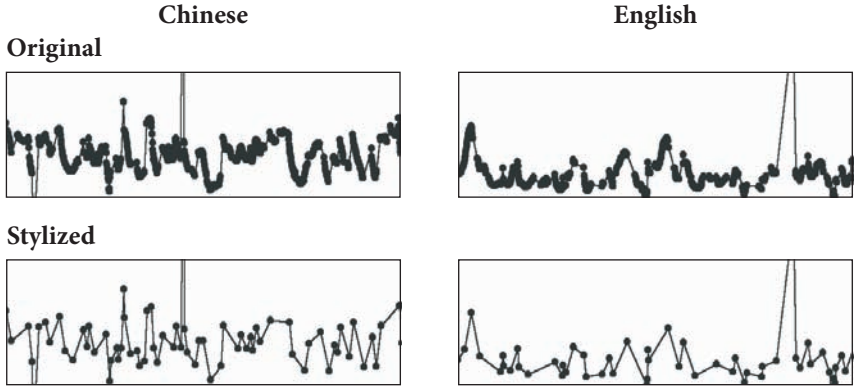

Figure 3. *Stylization of pitch.* X*-axis: Time;* Y*-axis: Pitch.*

larger values of F0 change (*y*-axis, positive or negative) indicate a quicker change of pitch (rising or falling). Note that extreme values of pitch change (e.g., greater than 1 or less than –1 in *y*-axis) may have been caused by pitch estimation errors.

Reprinted from *LACUS Forum XXIX: Linguistics and the Real World*, edited by Douglas W. Coleman, William J. Sullivan & Arle Lommel. 2003. Houston: LACUS.

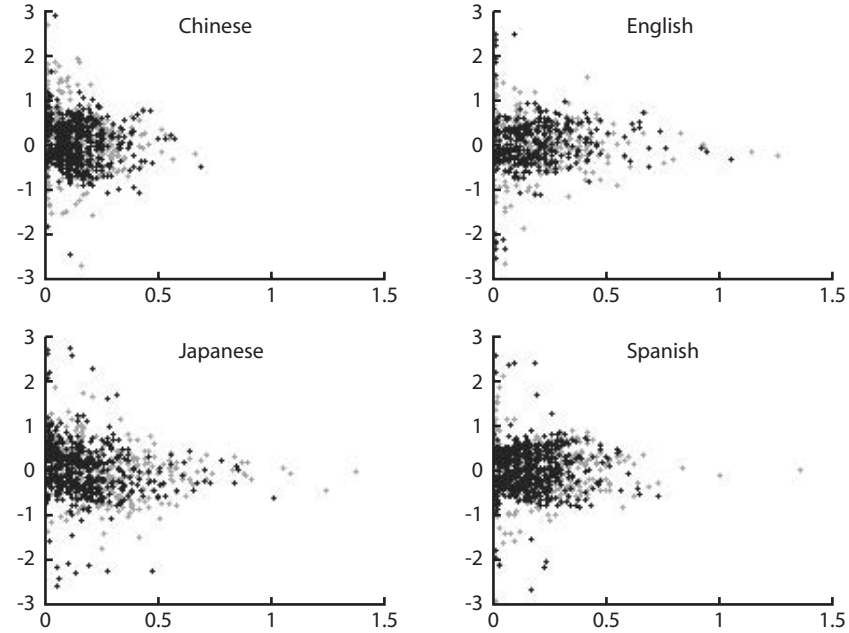

Figure 4. *Distribution of stylized pitch line segments.* x*-axis: Duration of line segments (sec);* y*-axis: Pitch change of line segments (log₂[Hz]). Darker dots indicate frames from male speakers; less dark, female speakers.*

Figure 4 shows that Chinese has many rapid changes in pitch. In contrast, English has many slow changes. Japanese shows a characteristic between Chinese and English. Spanish seems similar to English, but may have more quick and small changes.

3.2. INTENSITY AND HARMONICS-TO-NOISE RATIO.

3.2.1. MEAN AND STANDARD DEVIATION. Means and standard deviations of intensity and HNR were calculated for each language. The results did not show a clear difference among the languages. Japanese showed a slightly larger standard deviation in intensity, which was contrary to our expectation that Japanese would show a smaller standard deviation because it sounds more monotonic than the other languages. Such results suggest that the global analysis of intensity and HNR does not capture the prosodic difference.

3.2.2. INSTANTANEOUS INTENSITY OF HARMONICS-NOISE. Ramus et al. (1999) claim that languages of different rhythm types have different proportions of consonant and vowel durations in speech. Hence, we plotted the amplitudes of harmonics and noise to see the distribution of 'consonantal' and 'vocalic' frames (3 males and 3 females for each language). In Figure 5, dots higher and to the right in each graph indicate frames of greater intensity. Those located toward the top left indicate frames that have lower

Reprinted from *LACUS Forum XXIX: Linguistics and the Real World*, edited by Douglas W. Coleman, William J. Sullivan & Arle Lommel. 2003. Houston: LACUS.

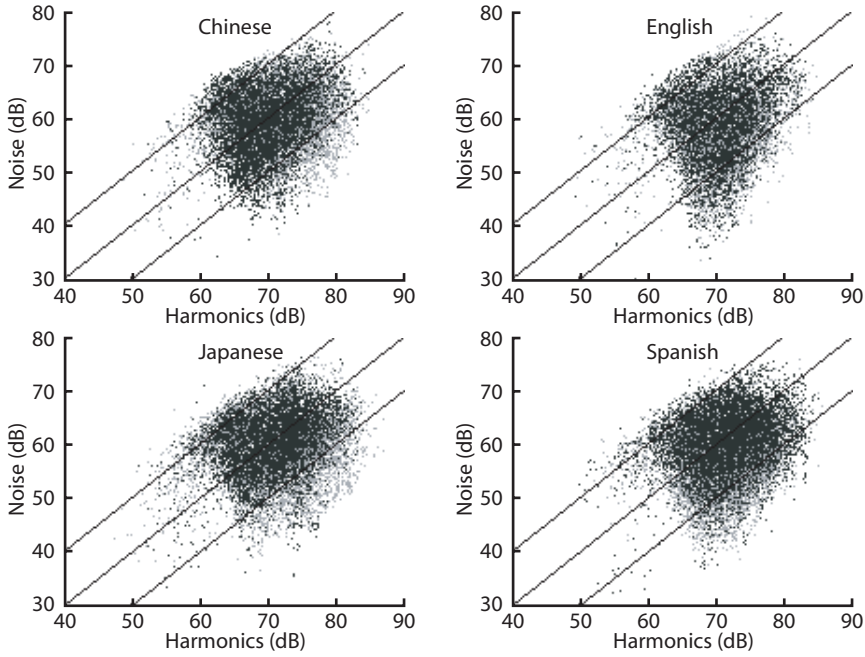

Figure 5. *Instantaneous intensity of harmonics-noise.* x*-axis: Amplitude of harmonics (dB);* y*-axis: Amplitude of noise (dB). Three auxiliary straight lines in each graph indicate HNR of 0, 10, 20 dB, respectively. Darker dots indicate frames from male speakers; less dark, female speakers.*

HNR (more consonantal), and those toward the bottom right indicate higher HNR (more vocalic). Three auxiliary straight lines in each graph indicate HNR of 0, 10, 20 dB, respectively.

These graphs clearly support the claim of Ramus et al. (1999) that English has the lowest proportion of vowel durations, Spanish has the next, and Japanese has the largest. The inverted triangle shape of the distribution of dots in the English graph clearly shows that English has a larger proportion of consonantal frames than vocalic ones. The Japanese graph shows that Japanese has a more equally balanced distribution of consonantal and vocalic frames. The shape of the distribution in Spanish falls between English and Japanese. The shape in Chinese, which is not analyzed in Ramus et al., may look similar to Japanese or Spanish, but it is not clear.

4. TIME ALIGNMENT OF LOCAL CHARACTERISTICS.

4.1. SMEARING OF LOCAL CHARACTERISTICS IN LONG-TERM ANALYSIS. In this section, we present an example indicating that a local characteristic is smeared in a long-term analysis, which shows the importance of the time alignment of characteristics in analysis.

Reprinted from *LACUS Forum XXIX: Linguistics and the Real World*, edited by Douglas W. Coleman, William J. Sullivan & Arle Lommel. 2003. Houston: LACUS.

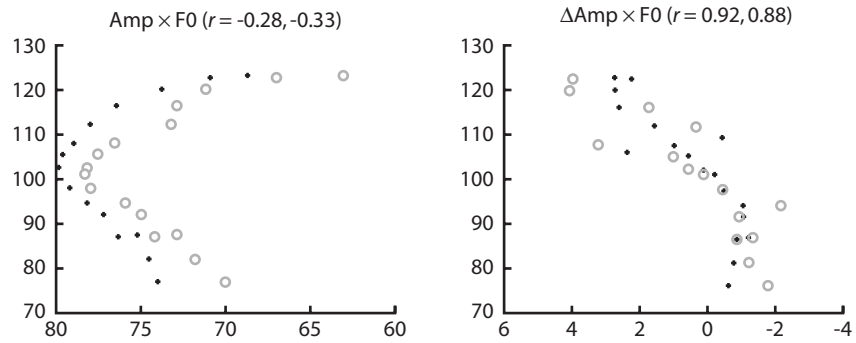

Figure 6. *Interaction of amplitude and pitch in isolated utterance of the word* pat. *Left panel:* x*-axis: Amplitude (dB);* y*-axis: Pitch (Hz). Right panel:* x*-axis: Differentiation of amplitude per 10 ms (dB);* y*-axis: Pitch (Hz). '✚' (black) indicates the interaction of intensity with pitch; '○' (gray), harmonics amplitude with pitch.*

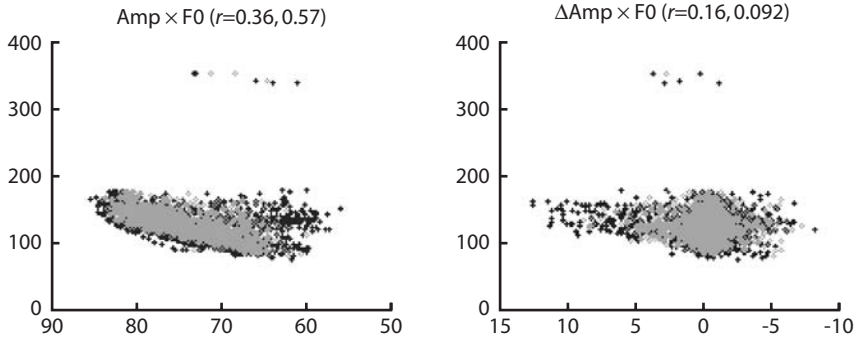

Figure 7. *Interaction of amplitude and pitch in continuous speech. Legends are the same as Figure 6.*

Figure 6 shows the interaction of intensity and pitch in the voiced section of the word *pat.* (Speech data is from our own recording.) In this example, intensity goes up from a low value, reaches a peak, and goes down, while pitch continuously goes down. If the amplitude is differentiated, it shows a strong correlation with pitch, such that the faster the amplitude goes up, the higher the pitch is; the faster the amplitude goes down, the lower the pitch is (see the right panel). The same characteristic was observed also in other words, such as *clap* and *splash*. This characteristic may be related to the nature of stress accent in English.

However, when the same analysis is applied to a whole continuous speech sample, such a characteristic is totally smeared, as shown in Figure 7. (Speech data are from Chino 1993.) To capture such syllable-wide characteristics, simple global statistical

Reprinted from *LACUS Forum XXIX: Linguistics and the Real World*, edited by Douglas W. Coleman, William J. Sullivan & Arle Lommel. 2003. Houston: LACUS.

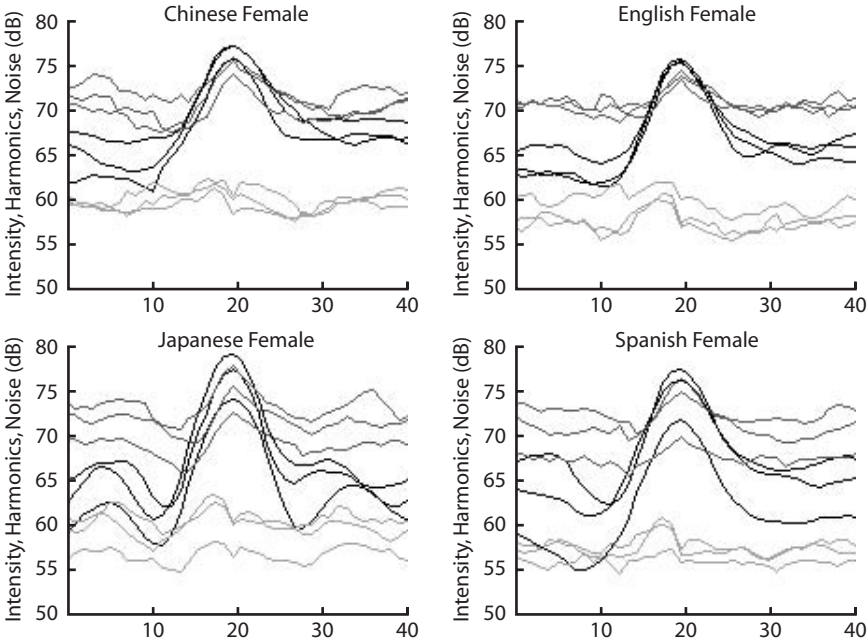

Figure 8. *Averaged syllables (Female; Intensity, harmonics amplitude, and noise amplitude).* x-*axis: Frame number (1 frame = 10 ms);* y-*axis: Amplitude (dB). The darkest lines indicate intensity contours; medium lines, harmonics amplitude; palest, noise amplitude.*

calculations are inadequate, and some operations for the time alignment of local characteristics are necessary.

4.2. AVERAGED SYLLABLES. Figure 8 shows averaged contours of intensity, harmonics amplitude, and noise amplitude (dark lines, medium lines, pale lines, respectively), which show the averaged shapes of syllables. Each graph shows averaged contours from three speakers (Hence, there are three lines for each type of contour).

These contours were calculated in the following way: First, an automatic algorithm selected from the speech signal 400 ms sections (40 frames) whose center is a local peak of the contour of harmonics amplitude. Next, the contours of intensity, harmonics amplitude, and noise amplitude were averaged across these sections. The result is that these contours are time aligned at the peaks of harmonics amplitude that can be regarded as syllable centers.

A universal characteristic of syllables is observed in these contours. See the Chinese graph (top left panel) for illustration. In the *x*-axis, 20 (the 20th frame) is the center of the averaged syllables. In the region of 10–20 in the *x*-axis, intensity is smaller and noise is greater than in 20–30 in the *x*-axis. This indicates that the syllable onset tends to have more consonantal elements, and that the syllable coda tends to

Reprinted from *LACUS Forum XXIX: Linguistics and the Real World*, edited by Douglas W. Coleman, William J. Sullivan & Arle Lommel. 2003. Houston: LACUS.

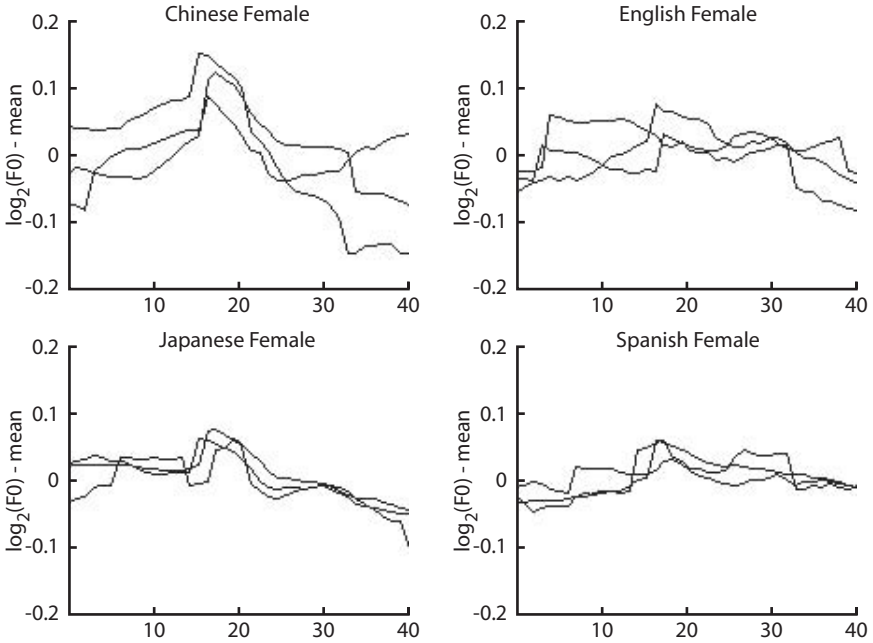

Figure 9. *Averaged syllables (Female; Pitch).* x-*axis: Frame number (1 frame = 10 ms);* y-*axis: Normalized pitch (log₂[Hz]).*

have more vocalic elements, such as the ending part of diphthongs, nasals, etc. Such a tendency can be observed in other languages, too.

Such representation also shows cross-linguistic differences. In Japanese, clearly there are bumps before and after the center peak (around the 5th and 35th frames). This indicates that Japanese syllables occur fairly regularly in time. Such bumps are also observed in Spanish, although they are less clear. (They are not well seen in Figure 8, but are clearer in a graph of males that is not included due to space limitations.)

Figure 9 shows the averaged contours of pitch. Chinese shows a tendency to large pitch falling within a syllable, and Japanese shows a smaller falling. English does not show such a tendency. In Spanish, it may be even considered that pitch tends to rise across consecutive syllables in some speakers. Such cross-linguistic difference is congruent with the asymmetry of positive and negative values of RCFF mentioned in Section 3.1.2.

5. CONCLUDING REMARKS. First, we proposed that pitch, intensity, and HNR are acoustic elements related to prosodic types.

Next, we have extracted local characteristics (histograms of RCFF, stylized line segments of pitch, and instantaneous intensity of harmonics-noise) as well as global characteristics (mean, standard deviation, and others of Eady's parameters). We proposed methods to analyze these local characteristics. We have shown the importance

Reprinted from *LACUS Forum XXIX: Linguistics and the Real World*, edited by
Douglas W. Coleman, William J. Sullivan & Arle Lommel. 2003. Houston: LACUS.

of local characteristics, or the 'distribution of parts', while indicating that some global characteristics, such as RCFF, may be useful as well. Finally, we proposed averaged syllables, which show the cross-linguistic difference of 'partial shapes' of acoustic elements as well as universal characteristics.

We have therefore extended Eady's F0 analysis on stress and tone languages to other prosodic types and proposed a more acoustic-natured measure than Ramus et al.'s rhythm analysis based on the phonological segmentation. We also introduced methods of analysis for local prosodic features; although we have not extensively discussed it in this paper, the literature suggests their importance (see Thymé-Gobbel & Hutchins 1996, etc.).

[1] We thank John T. Hogan and Lois M. Stanford for their comments. This research was partially supported by Gakunai Kyodo Kenkyu of Sophia University (Akira Ishikawa 2000–2002).

REFERENCES

BOERSMA, PAUL & DAVID WEENINK. 2002. *Praat,* version 4.0.18 (software). http://www.praat.org. (Accessed on June 6, 2002).

CHINO, EIICHI (ed.). 1993. *Sekai kotoba no tabi [Travel through the world's languages].* Tokyo: Kenkyu-sha. (Audio CDs).

EADY, STEPHEN J. 1982. Differences in the F0 patterns of speech: Tone language versus stress language. *Language and speech* 25:29–42.

ITAHASHI, SHUICHI (ed.). 2002. *Multilingual speech corpus (Report of the special research project for the typological investigation of languages and cultures of the East and West, Supplement).* University of Tsukuba, Japan. (CD).

RAMUS, FRANCK, MARINA NESPOR & JACQUES MEHLER. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition* 73:265–92.

THYMÉ-GOBBEL, ANN E. & SANDRA E. HUTCHINS. 1996. On using prosodic cues in automatic language identification. *Proceedings of International Conference on Spoken Language Processing '96*, 1768–71.

Reprinted from *LACUS Forum XXIX: Linguistics and the Real World*, edited by Douglas W. Coleman, William J. Sullivan & Arle Lommel. 2003. Houston: LACUS.