Interspeech (ISCA), pp. 1280-1283, 2004

Perceptual Discrimination of Prosodic Types and Their Preliminary Acoustic Analysis

Masahiko Komatsu[1], *Takayuki Arai*[2], *& Tsutomu Sugawara*[3]

[1]School of Psychological Science, Health Sciences University of Hokkaido, Sapporo, Japan
[2]Department of Electrical and Electronics Engineering, Sophia University, Tokyo, Japan
[3]Faculty of Foreign Studies, Sophia University, Tokyo, Japan
koma2@hoku-iryo-u.ac.jp, {arai, sugawara}@sophia.ac.jp

Abstract

A perceptual discrimination test was conducted to investigate whether humans can discriminate prosodic types solely based on suprasegmental acoustic cues. Excerpts from Chinese, English, Spanish, and Japanese, differing in lexical accent types and rhythm types, were used. From these excerpts, "source" signals of the source-filter model, differing in F0, intensity, and HNR, were created and used in a perceptual experiment. In general, the results indicated that humans can discriminate these prosodic types and that the discrimination is easier if more acoustic information is available. Further, the results showed that languages with similar rhythm types are difficult to discriminate (i.e., Chinese-English, English-Spanish, and Spanish-Japanese). As to accent types, tonal/non-tonal contrast was easy to detect. We also conducted a preliminary acoustic analysis of the experimental stimuli and found that quick F0 fluctuations in Chinese contribute to the perceptual discrimination of tonal/non-tonal accents.

1. Introduction

It is known that humans can discriminate languages based on prosodic cues to some extent. A number of perceptual experiments have been conducted to investigate whether humans can identify or discriminate languages or dialects by hearing real speech sounds or processed/synthesized sounds that simulate the prosody of speech (see [1] for stimulus types). Although these experiments suggest that prosody plays a role in language discrimination, they have been conducted rather sporadically. It is not clear yet whether humans can perceptually discriminate various prosodic types, such as lexical accent types and rhythm types that linguists have referred to, and how they are related to the acoustic properties of speech. To investigate these questions, it is necessary to conduct perceptual experiments with the acoustic cues parameterized.

Our study investigated whether humans can discriminate lexical accent types (tone, pitch, and stress accents) and rhythm types (stress-, syllable-, and mora-timed) and discussed how they are related to acoustic properties.

In the perceptual experiments, we used the "source" of the source-filter model as the stimuli. In their synthesis process, we controlled F0, intensity, and Harmonics-to-Noise Ratio (HNR). The temporal patterns of F0 and intensity are undoubtedly related to prosody. Besides, the source is related to the sonority feature (broad classification of phonemes) that seems to be an important contributor to rhythm. Recent studies on rhythm, such as [2-4], are based on the durations of consonant and vowel intervals, which means the acoustic properties that discriminate such phoneme classes are relevant. Ramus et al. ([3], p. 271 fn) wrote "[their] hypothesis should ultimately be formulated in more general terms, e.g. in terms of highs and lows in a universal sonority curve" rather than the consonant-vowel distinction. The source of the source-filter model significantly contributes to the perception of the sonority feature [1, 5].

The present paper is an extension of our previous study [6]. We report the results obtained from a larger number of participants in the same perceptual experiment as in our previous paper. Also, we conducted a preliminary acoustic analysis of the stimulus signals used in the perceptual experiment.

2. Speech data

We chose Chinese, English, Japanese, and Spanish as the languages representing prosodic types. Fig. 1 shows a provisional schematic layout of their lexical accent and rhythm types. In the figure, Chinese is situated at "tone accent" and "stress rhythm" tentatively because it is said to have both lexical tones and stress [7] although it is not traditionally classified in terms of rhythm [4].

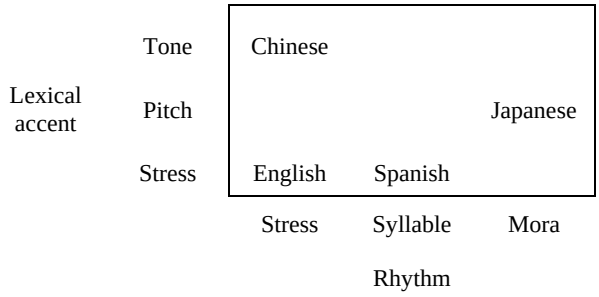

Figure 1: *Provisional schematic layout of prosodic types.*

The speech samples of English and Spanish were drawn from the MULTEXT prosodic database [8, 9]. The corpus consists of the recordings of 40 different passages for each of English, French, Italian, German, and Spanish. The same passages are translated into each language. The translation is rather free, and the expressions such as proper names are often changed to be adapted to the local culture. Besides the speech recording, the corpus contains the data of the stylized F0 curves created by the MOMEL algorithm [10], which extracts the macroprosodic component from the original F0.

Japanese samples were taken from the Japanese MULTEXT β [11, 12]. It has the "play" and "read" versions, the latter of which was used in our experiment. This corpus

also contains the MOMEL curves, but the data for one speaker were missing in the corpus, and we created substitutes for these.

We recorded the Chinese samples ourselves. The passages were translated into Mandarin by a Chinese collaborator. The recording was conducted in a soundproof studio (microphone: Sony ECM-MS957; DAT recorder Sony TCD-D100). Later, the audio data were down-sampled to 22 kHz (Onkyo SE-U77). When the F0 contours were stylized, the MOMEL algorithm was slightly modified (see [6]).

In the present experiment, for each language, 9 passages read by 3 speakers were used (Passage IDs *o6-o8* by Speaker 1, *p1-p3* by Speaker 2, and *p6-p8* by Speaker 3). The same passages were used across the languages to avoid emotional or attitudinal differences. Passages read by female speakers were selected because their pitch range is wider than that of males, and we expected the prosodic differences to be more distinct. The experimental stimuli were made from the first 5 seconds of these selected passages.

3. Experimental procedure

3.1. Signal processing

Six types of stimuli were created from the original speech. They simulated some characteristics of the original speech as described in Table 1. These sets can be grouped into three: those carrying amplitude information (Sets 1-3), the one carrying pitch information (Set 4), and those carrying both of them (Sets 5-6).

Table 1: *Stimulus sets. The middle column indicates what the stimuli simulate, and the right column indicates what they are made of.*

	simulates	is made of
Set 1	Intensity	white noise
Set 2	Intensity	pulse train
Set 3	Intensity, HNR	white noise and pulse train
Set 4	F0	pulse train
Set 5	Intensity, F0	pulse train
Set 6	Intensity, HNR, F0	white noise and pulse train

Set 1 is made of white noise.

Set 2 is made from a pulse train, whose F0 was constantly set to the mean value of the MOMEL curve.

Set 3 is a mixture of white noise and a pulse train. The amplitude contours 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 contour of harmonics, white noise was made based on the amplitude contour of noise, and they were added together. F0 of the pulse train was constant as well as in Set 2. In sum, voiced intervals in the original signal were represented as close to pulse trains; and unvoiced intervals, as close to white noise.

Set 4 is the pulse train created from the MOMEL curve. All unvoiced intervals were interpolated by MOMEL. The intensity was set constant.

Set 5 is the same as Set 4 except that it simulated the intensity of the original signal.

Set 6 is the mixture of white noise and a pulse train. It is the same as Set 3 except that it carried the stylized F0 contour. Note that unvoiced intervals such as [s] in the original signal did not carry F0 in Set 6 because they were converted to white noise, while they did carry interpolated F0 in Sets 4 and 5 because the whole signal was made of a pulse train.

All of these sounds were created with Praat (Version 4.1.6). They were created at the sampling frequency of 16 kHz. If amplitude less than a certain threshold continued more than 200 ms in the original signal, such intervals were regarded as pauses and suppressed to silence in the synthesized signal. Finally, they were tilted by −6 dB/oct to make them sound like a human voice.

Spanish data carried an unfavorable noise (seemingly a hum noise). Before the processing above, a noise reduction process (CoolEdit 2000) was applied only to the Spanish data.

We used the MOMEL data in the corpus if it existed. We calculated the MOMEL curve for one passage of Japanese which was accidentally missing from the corpus. We also calculated such curves for the Chinese speech. The Praat script (by G. Rolland, 2000; revised by S. Werner, 2002) was used for calculation, but the modification was necessary for Chinese (see [6]).

3.2. Perceptual experiment

Twenty graduate students and researchers (age: 22-45) specializing in linguistics, speech therapy, or speech engineering, voluntarily participated in the experiment. We asked those who were experienced in listening to various speech sounds because it was expected that the task would have been difficult for non-specialists.

The experiment was conducted in a soundproof studio. Stimuli were provided from a personal computer through a digital audio processor (Onkyo SE-U77) and headphones (Audio-Technica TH-65). Participants were allowed to adjust the volume of the stimuli according to their taste. The experiment was done with Praat.

Before the test session, sample sounds were given to each participant for a demonstration. The samples were, for each language, one original sound and Set 1-6 sounds created from it. The participant was asked to listen to all original sounds and at least one from each of Set 1-6 sounds. Then, the participant went through a short training session to become familiar with the operation of the program.

In the test session, the participant was asked to listen to a language pair and judge in which order the languages were presented. For example, after clicking the mouse, the Set 1 sound of Chinese and the Set 1 sound of English were successively played, and the participant clicked one of the two alternatives on the screen, namely, "(1) Chinese – (2) English" and "(1) English – (2) Chinese". The pairs were made so that they have the same passage ID, e.g. Chinese *o6* and English *o6*.

The test session was conducted from Set 1 to Set 6. Each set consisted of 6 subsets: Chi-Eng, Jpn-Spa Chi-Jpn, Eng-Spa, Chi-Spa, and Eng-Jpn. Each subset consisted of 6 trials (3 passage pairs [3 passages by different speakers, e.g. *o6*, *p1*, and *p6*] × 2 presentation orders [e.g. Chinese-English and English-Chinese]). The order of subsets within a set and the order of trials within a subset were changed for each participant. In total, the test session had 216 trials (6 trials × 6 subsets × 6 sets) and continued for about 1 hour.

4. Perceptual results: Correct response rates

In general, as the available information increases, the rates of correct responses increase, that is, discrimination gets easier. Correct response rates averaged across all languages are

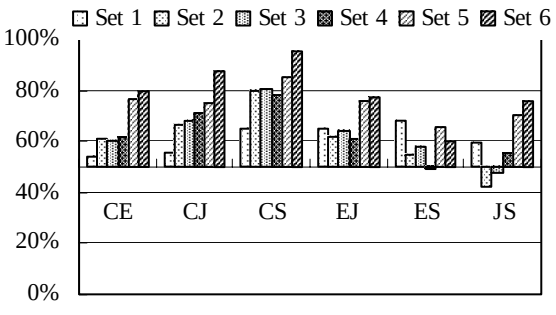

Figure 2: *Correct response rates for each language pair. "CE" stands for Chinese-English, etc.*

61.3%, 61.1%, 63.1%, 62.8%, 74.7%, and 79.3% from Sets 1 to 6. The rates for Sets 1-3, those with only amplitude information, are comparatively low, and the rate for Set 4, the one with only F0 information, is also low, but when such information combines (Sets 5-6) the rates get higher. This result is quite straightforward.

Looking into the correct response rates of each language pair, the situations get complicated. Fig. 2 shows the correct response rates for each language pair.

The Chinese-Japanese pair (CJ) and the Chinese-Spanish pair (CS) have good scores; they are easy to discriminate. This is understandable because these pairs are different both in lexical accent type (tonal/non-tonal contrast) and rhythm type.

The Chinese-English (CE) and English-Japanese (EJ) pairs (EJ) are not as easy to discriminate. Although they have comparatively good scores when both amplitude and F0 information are given (Sets 5-6), their scores are low when either of amplitude or F0 alone is given (Sets 1-4). The Chinese-English pair has the tonal/non-tonal contrast in accent type, but the fact that both languages have the stress may have an adverse effect. English and Japanese are different both in accent type and rhythm type, but the contrast of pitch and stress accents may not be as apparent as the tonal/non-tonal contrast.

The English-Spanish (ES) and Japanese-Spanish (JS) pairs are difficult to discriminate. Sometimes the rates go down below the chance level, which may suggest that the listeners were quite confused with these stimulus pairs. Considering the rhythm continuum [3], it is reasonable to suggest that English-Spanish and Japanese-Spanish are more difficult to discriminate than English-Japanese.

Considering all these together, rhythm types seem to be important in determining the discrimination difficulty, and the tonal/non-tonal contrast of accent type also seems relevant.

It is interesting to point out the tendency that in the language pair where the amplitude contributes, also F0 contributes. Simple conclusions such as the amplitude is related to rhythm type or F0 is related to accent type cannot be drawn.

Finally, consider the effects of HNR. The difference between Sets 1 and 3 and the difference between Sets 5 and 6 is the presence or absence of HNR information. While Set 3 shows higher rates than Set 1 only in 3 language pairs (CE, CJ, CS), Set 6 shows higher rates in 5 language pairs (all except ES). That is, the same information brought more improvement when combined with F0. It is inferred that it was easier in Set 6 to capture the timing relation of F0 change with the occurrence of some units such as syllables (e.g., whether or not an F0 change is within a syllable) than in Set 5 (and also Set 4) where F0 were interpolated during unvoiced intervals.

5. Acoustic analysis of stimulus signals

Several acoustic analyses (following [13, 14]) were conducted. Figure 3 shows the distribution of time interval and F0 movement between two successive MOMEL target points. It includes all target points in the stimulus signal used in the experiment (for each language, 5 sec × 9 passages = 45 sec). Table 2 shows the distribution of dots plotted in Fig. 3. (Chinese data may have been affected by our modification of the MOMEL algorithm in the experimental procedure.)

The means of time intervals indicate Chinese has quick F0 fluctuations, Japanese has smooth F0 movements, and English and Spanish fall between. The ranges are narrowest in Chinese (always quick) and widest in English (sometimes quick and sometimes slow).

The means of F0 movement, ascending or descending, indicate that there are smaller movements in Spanish and Chinese and larger movements in English and Japanese.

Figure 4 shows the distribution of intensity of harmonics and noise of every 10 ms analysis frame. This representation is expected to be related to the durations of consonantal and vocalic intervals in the speech signal, which in turn is related to rhythm. Japanese has the roundish distribution (balanced distribution), as pointed out in [13, 14], and the other languages are less balanced.

Figure 5 shows the averaged intensity contours time-aligned at their local peaks. From the intensity contour of each stimulus signal, 800 ms (80 frames) intervals centering the local peaks of the contour were cut out, and averaged. Thus, there are 9 averaged intensity contours (made from 9 stimuli) for each language. Of the four languages, Japanese shows the most regular pattern of the temporal change of intensity.

Acoustic peculiarities described above are partially related to perceptual scores. Chinese is distinct in having the quickest F0 fluctuations, and it is inferred that this enabled the perceptual tonal/non-tonal discrimination. (See Set 4 in CJ and CS; CE is adversely affected by stress.) In the intensity pattern, Japanese is distinct, but it is not the case in the perceptual scores in Sets 1-3. However, the combination of these acoustic characteristics must have played a role in perception.

6. Conclusions

Averaged correct identification rates produced results that conform to a rather common-sense view. Humans can discriminate lexical accent types and rhythm types. The more acoustic cues that are available, the easier discrimination is. The inspection of individual scores of language pairs suggested the importance of rhythm types and tonal/non-tonal contrast in lexical accent type, which supports the linguistic categorization of prosodic types.

Preliminary acoustic analysis of stimulus signals revealed difference among prosodic types. Quick F0 fluctuations in Chinese seem to contribute to the perceptual discrimination of tonal/non-tonal accent types. We have not yet succeeded in relating other individual acoustic characteristics of signals to perceptual discriminability. It seems that several acoustic characteristics in combination play a role in perceptual discrimination.

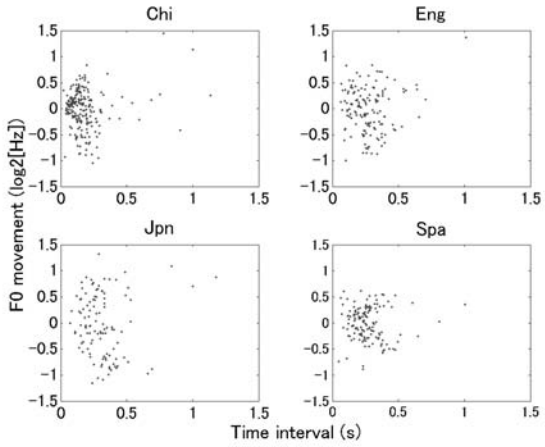

Figure 3: *Distribution of time interval and F0 movement between MOMEL target points.*

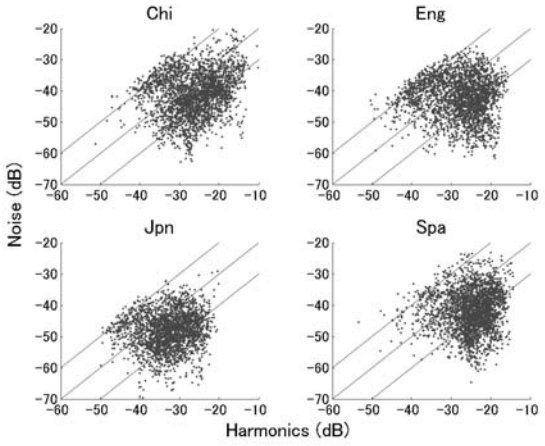

Figure 4: *Distribution of instantaneous intensity of harmonics and noise. Analysis frames lower than the global maximum amplitude by 20 dB or more are excluded.*

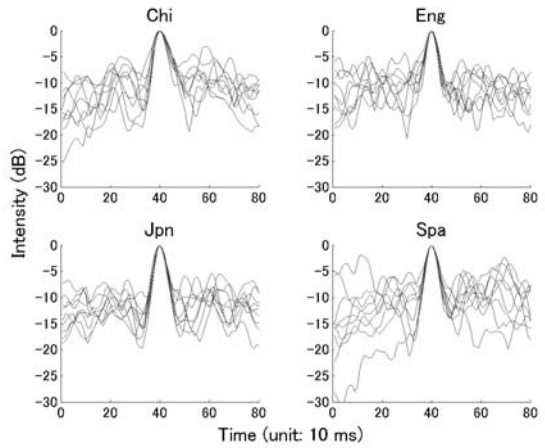

Figure 5: *Averaged intensity contours time-aligned at their local peaks.*

Table 2: *Distribution of time interval and F0 movement between MOMEL target points. Means and ranges exclude the highest 5% and the lowest 5% samples. Means of F0 movement are the means of absolute values.*

	N	Time interval (s)		F0 movement ($\log_2$[Hz])	
		Mean	Range	Mean	Range
Chi	218	0.15	0.33	0.26	1.23
Eng	141	0.26	0.44	0.34	1.41
Jpn	110	0.29	0.39	0.49	1.75
Spa	142	0.25	0.37	0.25	1.02

7. References

[1] Komatsu, M. (2002). What constitutes acoustic evidence of prosody? The use of LPC residual signal in perceptual language identification. *LACUS Forum, 28*, 277-286.

[2] Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *J. Acoust. Soc. Am., 105*, 512-521.

[3] Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73*, 265-292.

[4] Grabe, E., & Low, E. L. (2002). Durational variability in speech and the Rhythm Class Hypothesis. In C. Gussenhoven & N. Warner (Eds.), *Laboratory Phonology 7* (pp. 515-546). Berlin: Mouton de Gruyter.

[5] Komatsu, M., Tokuma, S., Tokuma, W., & Arai, T. (2002). Multi-dimensional analysis of sonority: Perception, acoustics, and phonology. *Proc. ICSLP 2002*, 2293-2296.

[6] Komatsu, M., Arai, T., & Sugawara, T. (2004). Perceptual Discrimination of Prosodic Types. *Proc. Speech Prosody 2004*, 725-728.

[7] Hirst, D., & Di Cristo, A. (1998). A survey of intonation systems. In D. Hirst & A. Di Cristo (Eds.), *Intonation Systems: A Survey of Twenty Languages* (pp. 1-44). Cambridge, UK: Cambridge University Press.

[8] Campione, E., & Véronis, J. (1998). A multilingual prosodic database. *Proc. ICSLP '98*, 3163-3166.

[9] Campione, E. (Ed.). (1998). *MULTEXT prosodic database* [CD-ROM]. Paris: European Language Resources Association.

[10] Hirst, D., Di Cristo, A., & Espesser, R. (2000). Levels of representation and levels of analysis for the description of intonation systems. In M. Horne (Ed.), *Prosody: Theory and Experiment* (pp. 51-87). Dordrecht, The Netherlands: Kluwer Academic.

[11] Kitazawa, S., Kitamura, T., & Itoh, T. (2002). Nihongo MULTEXT ni okeru inritsu joho no bunseki to shuroku. In *Proc. 2001 2nd Plenary Meeting and Symposium on Prosody and Speech Processing, Tokyo* (pp. 39-50).

[12] Kitazawa, S. (Ed.). (2002). *Japanese MULTEXT* (β version) [CD-ROM]. Shizuoka University, Japan.

[13] Komatsu, M., & Arai, T. (2003). Acoustic realization of prosodic types: Constructing average syllables. *LACUS Forum, 29*, 259-269.

[14] Komatsu, M., & Miyakoda, H. (2003, August). *Acoustic measurement of rhythm types: A stress language vs. a mora language.* Paper presented at the 38th Linguistics Colloquium, Péter Pázmány Catholic University, Piliscsaba, Hungary.

* We thank Lois M. Stanford for her comments on the draft.