

Analysis of impression-prosody mapping in communicative speech consisting of multiple lexicons with different impressions

Yoko Greenberg¹, Hiroaki Kato², Minoru Tsuzaki³, Yoshinori Sagisaka¹

¹ *GITI/Language and Speech Science Res. Lab Waseda University, Japan*

², *NICT/ATR Media Information Science Lab, Japan*, ³ *Kyoko City University of Arts, Japan*

yoko.kokenawa@toki.waseda.jp, kato.hiroaki@nict.go.jp, minoru.tsuzaki@kuca.ac.jp, ysagisaka@gmail.com

Abstract

Aiming at prosody control for communicative speech synthesis, we analyzed communicative prosody of phrases consisting of adverbs, adjectives and particles showing multiple impression combinations in six directions of three-dimensional impression space (confident-doubtful allowable-unacceptable and positive-negative) derived by dimension reduction using Multi Dimensional Scaling. Through this analysis, we tried to generalize our impression-prosody mapping scheme previously proposed for phrases with single impression. Analysis showed that changes in average F0 height, F0 dynamics and utterance duration were systematically explained as a sum of single control characteristics assigned by each impression using the impression-prosody mapping scheme. These results suggest the applicability of our impression-prosody mapping scheme to more general inputs consisting of multiple lexicons with different word impressions.

1. Introduction

High quality speech output from corpus-based text-to-speech synthesis (TTS) systems has enhanced the wide use of synthetic speech not only for simple reading purposes but also for interactive applications where communicative functions of speech are expected. In these applications, traditional TTS reading-style speech is not sufficient regardless of how high is the speech quality as reading-style speech. In communications, we use speech not only to convey so-called linguistic information but also other communicative information that written language cannot express. This speech specific communicative

information has not yet been well studied but is of great use for smooth information exchange between humans. We need a new scheme to specify this communicative information and find its correlation to speech prosody to synthesize human friendly speech.

In speech synthesis, *non-reading* prosody variations have been extensively studied for specific characteristics assigned externally such as emotional speech [1]-[3]. In most of these studies, additional control factors such as emotion-types have been pre-assigned and added to conventional prosody characteristics. Though these studies can contribute to generate non-reading speech, they only take care of overall pre-specified categorical control as seen in emotional speech. They cannot cope with utterance specific control. As the specification of utterance dependent communicative prosody by itself is quite difficult, we have been approaching the utterance specific control by employing correlations between impression of lexicon and communicative prosody [4]-[8]. Instead of specifying prosody type as prototypical single category, we have specified communicative prosody using word attributes of constituents.

Starting from simple sentences consisting of an adverb and an adjective, we confirmed that we can generate communicative prosody of the utterances consisting of adjectives showing impressions such as positive and negative preceded by the adverbs showing magnitude [4]. In the proposed prosody generation scheme, the communicative prosody could be obtained as a sum of a conventional prosody component and a communicative new one predicted from word impressions of constituting lexicons [6]-[8]. Based on this lexicon impression-driven communicative prosody generation (hereafter we call it *impression-prosody mapping*), we have successfully generated the

communicative prosody for single lexicon phrases with six prototypical impressions out of three dimensional impressions.

To scale up this impression-prosody mapping to more general sentences such as “Quite interesting, isn’t it?” and “Not so dirty”, we need to know how the communicative prosody behaves when multiple impressions meet together. In this paper, we analyzed communicative prosody characteristics of utterances consisting of multiple lexicons with different word impressions to check if their prosody can be consistently explained by summing up prosodic characteristics derived from each impression-prosody mapping. If we can confirm these additive control characteristics, it will be possible for us to employ a simple additive computational model for prosody generation to these general utterances.

In the following sections, first we introduce an impression-prosody mapping, a communicative prosody generation scheme where a word impression of input is employed to specify default communicative prosody for the phrase with single lexicon [6]-[8]. Then, we design an utterance set to analyze communicative prosody characteristics for phrases consisting of multiple lexicons with different word impressions as well as the details of communicative speech collection of the selected phrases in Section 3. We show the analysis results and discuss in Section 4. Finally, we sum up our findings and state further works in Section 5.

2. Impression-prosody mapping for communicative prosody generation

In conversational situations, lexicons constituting an utterance by themselves seem to have strong correlation to its prosody. For example, when we say something confident, we generally employ lexicons relating to confidence and prosody as well as a default. That is, it is very likely that we say something confident with corresponding prosody style which can be predicted from constituent lexicons. Though in real situations, there are more complex factors relating to a speaker’s will or attitude. If we would like to have prosody reflecting these factors, they should be supplied additionally as extra-communication information. In this paper, we focus on default prosody generation driven from constituent lexicons as the first step towards communicative prosody generation.

By analyzing communicative prosody characteristics of one word utterances of “uhm” using impression metric and MDS analysis, we found that their freedoms of perceptual impression can be reduced to three-dimensional space showing *confident-doubtful*, *allowable-unacceptable* and *positive-negative* impressions [5]. Applying this three-dimensional impression expression to single lexicons, we found the following correlations between impressions of lexicons and their corresponding communicative prosody characteristics [6]. F0 average height gives the distinction between high group (*confident*, *allowable* and *positive*) and low group (*doubtful*, *unacceptable* and *negative*). F0 dynamics

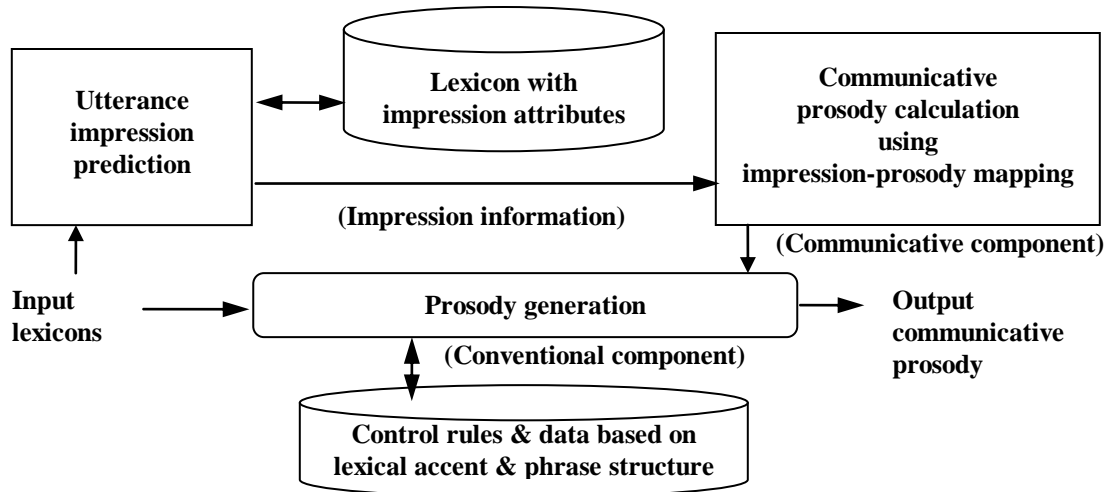


Figure 1 Communicative prosody generation using impression prediction by input lexicons

correlates to the impression distinctions of *doubtful-confident* and *unacceptable-allowable*. Total utterance duration also correlates to perceptual impressions. Longer durations are employed for utterances with *doubtful* or *unacceptable* impressions while the shorter ones are used for *confident* or *allowable* impressions.

Based on impression-prosody correlations for a phrase with a single lexicon, we proposed a communicative prosody generation scheme as shown in Figure 1 [6]. In this generation scheme, we add a

Table 1 Lexicons employed for test phrases to measure their effects on communicative prosody

(a) Adjectives expressing *positive/negative* impression

| positive | | negative | |
|----------|----------------------------------|-----------|----------------------------------|
| Japanese | corresponding English expression | Japanese | corresponding English expression |
| kirei | beautiful, clean | kitanai | dirty |
| umai | delicious | mazui | unsavory |
| yasasii | nice, kind | kibisii | strict |
| omosiroi | interesting | tumaranai | boring |

(b) Final particles expressing allowable/unacceptable and confident/doubtful impression

| degree | Japanese particle | corresponding English expression |
|----------------|-------------------|----------------------------------|
| confident | da | ! |
| less confident | yone | isn't it? |
| less doubtful | kana | wonder |
| doubtful | nano (ka) | ? |

| degree | Japanese particle | corresponding English expression |
|-------------------|-------------------|----------------------------------|
| allowable | yo | ! |
| less allowable | kamo | maybe |
| less unacceptable | naikamo | maybe not |
| unacceptable | nai | not |

(c) Adverbs expressing the degree

| degree | Japanese adverb | | corresponding English expression |
|-------------|-------------------------|----------------------|----------------------------------|
| | in affirmative sentence | in negative sentence | |
| strong | sugoku | chittomo | extremely |
| less strong | sootoo | sahodo | very |
| less weak | wariai | taishite | quite |
| weak | sokosoko | amari | relatively |

communicative component to a conventional reading-style one derived from a TTS system. For F0 generation, the conventional component and the communicative one are added in the parameters of F0 generation model such as the command-response model proposed by Fujisaki [9]. Using this impression-prosody mapping, we have already confirmed that communicative prosody for phrases consisting of single lexicon such as “Absolutely”, “Fishy”, “Agree”, “Unacceptable”, “Impossible”, “Interesting” and “Dirty” can be successfully obtained by modifying reading-style prosody based on a word impression [6]-[8].

To generalize this impression-prosody mapping, we have to understand how the word impression of each lexicon contributes to their output communicative prosody when a phrase consists of multiple words with different impressions. If communicative prosody could be obtained as the sum of each effect, we can simply apply the additive modeling framework to generate communicative prosody for more general phrases. To confirm this impression-prosody mapping characteristics, we analyzed the communicative prosody of utterance phrases consisting of multiple lexicons with different word impressions.

3. Design of phrases consisting of multiple lexicons with different word impressions

3.1. Phrase selection

To examine how each word impression of constituting lexicons respectively affects their communicative prosody, we analyzed the communicative prosody characteristics of the phrases consisting of multiple lexicons with different word impressions. The phrases were designed to include two lexicons reflecting different word impressions out of six prototypical impressions (*confident*, *doubtful*, *allowable*, *unacceptable*, *positive* and *negative*). We used four sets of adjectives for *positive-negative* contrast as shown in Table 1 (a). Four final particles of are employed to contrast *confident-doubtful* and the other four final particles are for *allowable-unacceptable* contrast as shown in Table 1 (b). In Japanese, as final particles convey modality in communication, they are frequently employed in daily conversations. As shown in the table, there exist quite a few particles giving the above impressions with different degrees.

To quantitatively analyze the communicative prosody changes depending on the word impressions of input lexicons, we also employed adverbs as shown in Table 1 (c). The degrees of *positive-negative*

contrast of adjectives were given by these adverbs. Though it is ideal to employ the same adverbs for all *confident-doubtful*, *allowable-unacceptable* impressions, certain adverbs are used only with affirmative sentence while others are linked with negative sentences. For this reason, we adopted different adverbs for affirmative sentences and negative ones by balancing the degree as shown in Table 1 (c). In total, we made 256 different phrases.

3.2. Communicative utterance collection

To analyze the communicative prosody of the designed phrases, the naturally spoken speech samples were first recorded. To obtain communicative speech samples that reflect daily conversations as closely as possible, we suggested the plausible situations associated with each utterance phrase to encourage the speakers to imagine the appropriate situation and to utter each phrase in a conversational situation. For example, if they are asked in nervous manner to identify another person's character as "Extremely nice!" In total, 896 speech samples were uttered by five (two males and three females) native Tokyo standard Japanese speakers whose ages were ranging 24-35. Among the 256 phrases, 64 phrases were presented as one set. All speakers were instructed to utter while comfortable, prior to any levels of fatigue. Each speaker uttered two sets in average. These samples were also recorded in a reading style after recording the communicative speech.

4. Communicative prosody analysis of phrases with multiple word impressions

4.1. Communicative prosody analysis

To analyze the prosody control differences in respect to the word impression of input lexicons, the communicative speech samples were compared with the reading style ones. To compare with our previous observations of the variations in one word utterance "uhm" [8], we checked the following three characteristics of impression-prosody mapping .

- (1) The *positive-negative* word impression and average F0 height
i.e. positive/negative impressions give higher/lower F0
- (2) The *confident-doubtful*, *allowable-unacceptable* word impressions and F0 dynamic pattern
i.e. confident-doubtful impressions are ordered as Fall, Rise&Fall, Gradual fall and Rise in F0 dynamics, *allowable-unacceptable* impressions are ordered as Fall, Gradual fall, Rise and Rise&Fall in F0 dynamics

- (3) The *confident-doubtful*, *allowable-unacceptable* word impressions and average utterance duration
i.e. (confident, allowable)/(doubtful, unacceptable) impressions give shorter/longer utterance duration

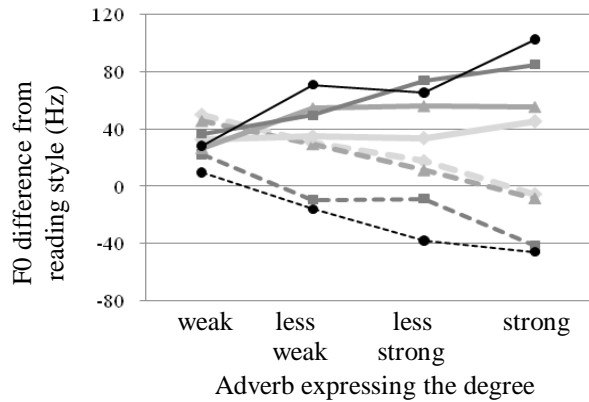
For the analyses, we measured the prosody parameter differences. As for the average F0 height and the utterance duration, the values in reading style speech samples were simply subtracted from the ones in communicative style. For the F0 dynamic patterns, the ten points of F0 height values of each communicative speech samples were extracted. They were classified into each of the four F0 dynamic patterns or non-applicable one by hand.

To see how the extracted prosody characteristic values were respectively related to the word impressions of input lexicons, the average of the F0 height difference values were calculated depending on the adverb showing the degree of the meaning enhancement of the following adjectives with *positive/negative* word impressions. As for the *confident-doubtful* and *allowable-unacceptable* word impressions contribution to the communicative prosody, the number of the F0 dynamic pattern classified into each of four F0 dynamic patterns and the average utterance duration values were calculated in relation to the degree of impressions indicated by the final particles with *confident-doubtful* and *allowable-unacceptable* impressions.

4.2. Communicative prosody characteristics based on word impressions of input lexicons

To check the impression-prosody characteristic (1) stated in the previous section, we measured F0 average differences between communicative speech and reading one. Figure 2 shows the average F0 differences for adjectives with *positive-negative* impression with an adverb showing degree. For both cases (a) and (b) with different final particles, the *positive-negative* word impressions steadily correspond to the output communicative prosody as shown in the up/down sloping curve depending on the degree of the *positive/negative* impressions overall. For the phrase including (a) *doubtful* (b) *unacceptable* final particles, the down trends were systematically shown as the phrase meaning turned reversed. These results support that *positive-negative* word impressions were consistently related to F0 average in these mixed conditions. The *positive/negative* impressions of adjectives were directly manifested as the higher/lower average F0 of communicative speech respectively. Moreover, the *confident-doubtful* and *allowable-unacceptable* impressions of the final

a) *Confident-doubtful* final particle phrases
 ● confident ■ less confident ▲ less doubtful ◆ doubtful



b) *Allowable-unacceptable* final particle phrases
 ● allowable ■ less allowable ▲ less unaccep. ◆ unaccep.

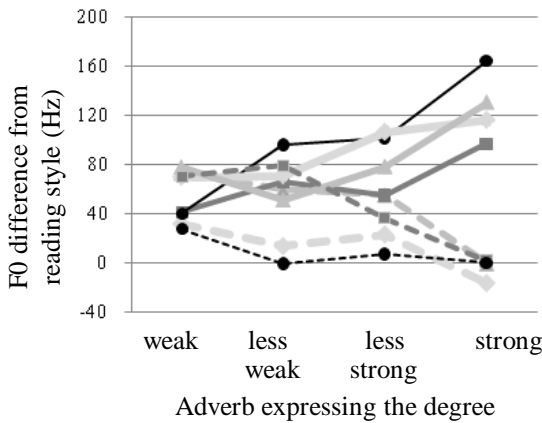
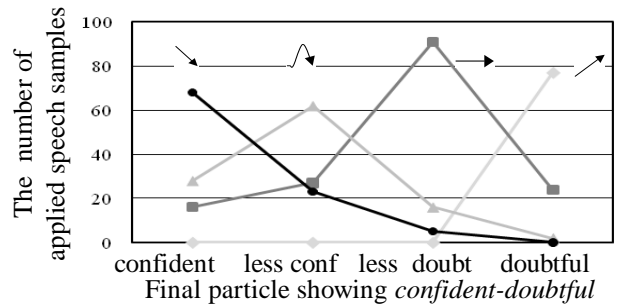


Figure 2 Average F0 difference changes with the degree of positive-negative word impressions (Solid/dotted lines correspond to utterances with positive/negative adjectives.)

particles systematically change the average F0 increase slopes.

To check the impression-prosody characteristic (2) the *confident-doubtful* and *allowable-unacceptable* impressions to F0 dynamic pattern in the previous section, we measured F0 dynamics differences by counting the number of four shapes (Fall, Gradual fall, Rise and Rise&Fall). Figure 3 shows the distribution differences of these four shapes for *confident-doubtful*, *allowable-unacceptable* impressions of final particles. As the figure shows, most samples categorized in each pattern appeared in the order of Fall, Rise&Fall, Gradual fall and Rise from *confident* impression to *doubtful* impression and Fall, Gradual fall, Rise and Rise&Fall from *allowable* impression to *unacceptable* impression.

a) *Confident-doubtful* final particle phrases
 ● Fall ■ Rise&Fall ▲ Gradual fall ◆ Rise



b) *Allowable-unacceptable* final particle phrases

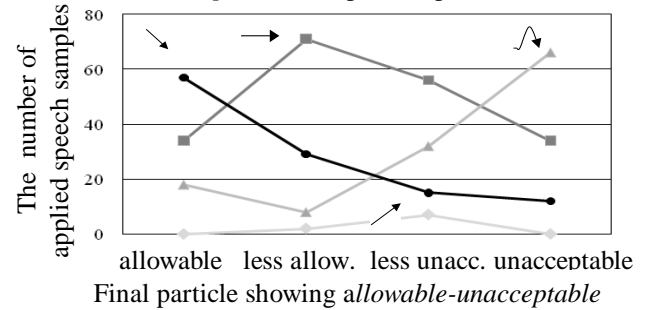
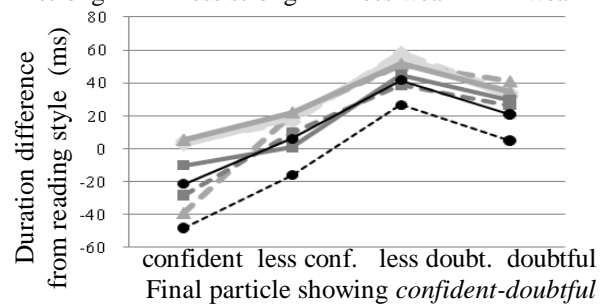


Figure 3 Average utterance duration changes with the degree of *confident-doubtful/allowable-unacceptable* word impressions

a) *Confident-doubtful* final particle phrases
 ● strong ■ less strong ▲ less weak ◆ weak



b) *Allowable-unacceptable* final particle phrases

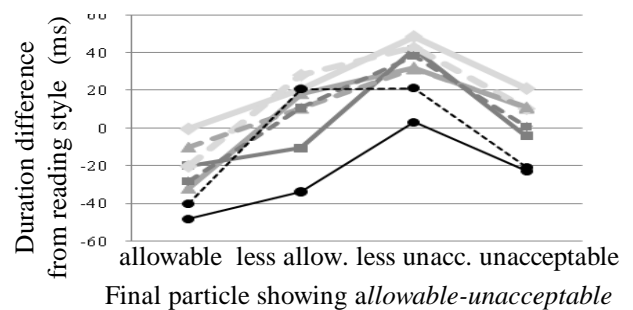


Figure 4 Average utterance duration changes with the degree of *confident-doubtful/allowable-unacceptable* word impressions

For the impression-prosody characteristic (3) in the previous section, the utterance duration differences relating to the degree of the *confident-doubtful/allowable-unacceptable* word impressions are shown in Figure 4. The values were disposed to compare based on the degree of the *confident-doubtful* or *allowable-unacceptable* word impressions with respect to each phrase with adverbs showing magnitude to be preceded to the adjectives with *positive-negative* word impressions. As shown in the figure, the utterance duration has consistently become shorter for the phrases including the final particles with more *confident* or *allowable* word impressions overall. That is, the *confident* or *allowable* / *doubtful* or *unacceptable* impression of input lexicons gives the shorter/ longer duration respectively.

For the phrases with extreme *doubtful* or *unacceptable* impressions, we found consistently shorter duration than phrases with *less doubtful* or *less unacceptable* impressions. We are speculating that these phenomena result from extra factors of the particles used in these experiments showing assertion which is slightly different from the lexicons such as “suspicious” or “wrong” which we originally intended as “doubtful” or “unacceptable” impressions. Though the control tendency is consistent, we need more careful reconsideration to the impression attributes.

The above results show that the impression-prosody mapping characteristics observed in single impression phrases consistently manifested in phrases with multiple words with impressions. These observations support the invariance of impression-prosody mapping characteristics and suggest the possibility of communicative prosody control by assembling each effect of multiple lexicons with different impressions.

5. Conclusions

To derive prosody characteristics for speech synthesis with communicative speech, we have analyzed the prosody of communicative utterances consisting of adverbs, adjectives and particles showing the six prototypical impressions (*confident, doubtful, allowable, unacceptable, positive* and *negative*). Through the analysis, we confirmed that the impression-prosody mapping characteristics found in phrases with single impression are also applicable to phrases with multiple impressions. Namely, the lexicons with *positive/negative* impressions consistently change the average F0 height while the lexicons with *confident-doubtful* or *allowable-*

unacceptable impressions systematically change F0 dynamic patterns of the communicative speech as well as the utterance duration. These results suggest the applicability of impression-prosody mapping approach for general communicative prosody generation by assembling each effect of multiple lexicons with different impressions. As same as other corpus-based synthesis scheme, we can think of computational modeling of communicative prosody generation along with impression-prosody mapping scheme as an immediate research topic.

Acknowledgments

This work was supported in part by the Grant-in-Aid for Scientific Research (B) No. 18300063 and 20300069, JSPS.

References

- [1] N. Audibert., D. Vincent, V. Auberge and O. Rosec, “Expressive speech synthesis: Evaluation of a voice quality centered coder on the different acoustic dimensions,” Proc. Speech Prosody 2006, 2006, pp. 525–528.
- [2] N. Campbell, A. Iida, F. Higuchi and M. Yasumura, “A corpus-based speech synthesis system with emotion,” Speech Communication 40, 2003, pp.161–187.
- [3] J. Tao, L. Xin and L. Yin, “Realistic visual speech synthesis based on hybrid concatenation method”, IEEE Trans, 17, 3, 2009, pp. 469-477.
- [4] Y. Sagisaka, T. Yamashita and Y. Kokenawa, “Generation and perception of F0 markedness for communicative speech synthesis” Speech Communication 46, 2005, pp. 376-384.
- [5] Y. Kokenawa, M. Tsuzaki, H. Kato and Y. Sagisaka, “F0 control characterization by perceptual impressions on speaking attitudes using Multiple Dimensional Scaling analysis”, Proc. ICASSP, Mar.2005, pp. 273- 276.
- [6] Y. Greenberg, M. Tsuzaki, H. Kato and Y. Sagisaka, “Communicative speech synthesis using constituent word attributes”, Proc. INTERSPEECH 2005, Sep.2005, pp. 517-520.
- [7] Y. Greenberg, M. Tsuzaki, H. Kato and Y. Sagisaka, “A trial of communicative prosody generation based on control characteristic of one word utterance observed in real conversational speech”, Proc. Speech Prosody 2006, May 2006, pp. 37-40.
- [8] Y. Greenberg, N. Shibuya, M. Tsuzaki, H. Kato and Y. Sagisaka, “Analysis on paralinguistic prosody control in perceptual impression space using multiple dimensional scaling”, Speech Communication 51, 2009, pp. 585–593.
- [9] H. Fujisaki and K. Hirose, “Analysis of voice fundamental frequency contours for declarative sentences of Japanese,” J. Acoust. Soc. Jpn. (E), 5, 1984, pp. 233–242.