Modeling prosody variations for communicative speech and the second language towards trans-disciplinary scientific understanding

Yoshinori Sagisaka

Global Information and Telecommunication Institute, Department of Applied Mathematics, Linguistics and Speech Science Research Laboratory of Waseda University, Tokyo

ysagisaka@gmail.com

Abstract

In this paper, our research studies on prosody variation modeling are introduced for communicative prosody characterization and the objective evaluation of the second language (L2) timing control characteristics. For communicative prosody characterization, a possibility of lexicon driven control and further needs of dialogue-act modeling are discussed. For the objective evaluation of L2 prosody, the possibility of scientific understanding of timing control characteristics and the needs of perceptual studies are demonstrated. Through the introduction of these studies, I would like to show the necessity and the merits of trans-disciplinary research collaboration among multiple research areas relating to speech science and technologies including linguistics, phonetics, speech science, information processing, and language education. Finally, the research efforts are introduced for international research consortium AESOP (Asian English Speech cOrpus Project) to collect commonly sharable learner's spoken language data and knowledge of L2 studies for trans-disciplinary scientific understanding.

Index Terms: communicative speech synthesis, second language learning, fundamental frequency control, timing control, para-linguistic information, prosody evaluation, prosody perception

1. Introduction

As speech prosody is an interdisciplinary research field, I guess that the readers of this article have quite wide academic background and research interests. Though I myself have been devoting most of my research life as a researcher in speech information processing field, I can imagine that research interests could be varied from topic to topic and the final goals and the values can be different from field to field. Whenever I started to work in a new field, I felt that my own views were not always identical to what had been mainly treated in the corresponding research field, which reminded me of new disciplines and different scope in the corresponding new fields. The more we study the topics in new different fields, the wider thinking we have and we will be free from field-intrinsic prefixed knowledge, unconsciously constrained thinking and routine approach.

In this paper, I would like to introduce two research topics relating to prosody that I have encountered during my research life. Through the introduction of the prosody related topics on communicative prosody and the second language studies, I want to point out the importance and usefulness of research from trans-disciplinary view points and research collaboration of researchers with different disciplines and values. In the following Section 2, the research topic on communicative prosody is introduced. Throughout this introduction, I would like to point out some possibility of the prosody characterization which has been given up as *para-linguistics* (*i.e.* out of linguistics) in traditional linguistics. In Section 3, our research efforts on the second language (L2) speech evaluation are introduced. The possibility of objective evaluation of L2 learner's timing characteristics is discussed. By assembling our knowledge on rhythm and timing studies, we found the necessity of research collaboration of corresponding research fields. This necessity let us work to form AESOP (Asian English Speech cOrpus Project) research consortium which is described in Section 4. Finally, I would like to wrap up my presentation by looking for further studies on computing prosody.

2. Studies on communicative prosody

2.1. Communicative prosody

In traditional linguistics, quite physiological topics and colloquial speech intrinsic phenomena have been excluded from the theoretical research targets [1]. This restriction of research targets has nicely worked to build many essential linguistic theories without being bothered by treating many irregular linguistic phenomena as performance. This restriction has also been inherited in prosody studies. As a natural result, most of research efforts in prosody have been devoted to the studies on speech prosody of well formed utterances and its linguistic structures or accentual properties which is closely related to the interests in written language. In this sense, mainly speech phenomena closely related to written language have been widely studied and others such as the variations among speech utterances have not been systematically studied at least from linguistic viewpoint. As even the description of them has not yet been discussed, we have difficulties to describe the existing problems.

On the other hand, recently, as exemplified by the number of published papers in the series of Speech Prosody Conferences, many research interests have been attracted to so called "*para-linguistics*". There are quite many paper submissions on emotional speech or other prosodic phenomena which have not been treated in traditional studies. As there has been little interest to specify the prosody variations observed in real communications and characterize their communicative functions, these phenomena have been simply treated as *para-linguistics* "out-of-linguistics", though it plays an essential role in speech communication. Are they really to be treated as *para-linguistics* "out-of-linguistics" ?

We have started to analyze non-reading speech aiming at synthesis of more realistic human speech. To distinguish with conventional reading-style speech which has been a research

target for a long time in text-to-speech technology, I would like to use the term "communicative prosody" hereafter to imply prosody which is generated by considering its communicative function. It can be considered as speech attributed some information by a sender to a receiver. To enable communicative speech output, if we can specify some of communicative characteristic in prosody domain, we will be able to add new communicative factors to prosody control by considering its function. As it looked quite difficult to characterize the prosody of communicative speech in general, we started the analysis of simple utterances consisting of an adjective with an adverb expressing magnitude [2].

2.2. Lexicons and communicative prosody

As the first step towards communicative prosody characterization, we started to find the differences of prosody between read speech and communicative one obtained from simulated conversations [2]. For speech contents, we designed two-phrase utterances consisting of Japanese adjective and adverb phrases expressing different degree under designed conversational situations. Five paired adjectives were employed to show either positive or negative meaning (e.g. beautiful/dirty, interesting/boring) and six adverbs expressing degrees were employed. Through both F0 observation of communicative speech with various F0 heights, we could have confirmed the following characteristics.

- There were consistent F0 differences between the communicative speech prosody and the read one depending on adjective's attribute and degree of adverb
- Positive/negative adjectives respectively give average F0 increase/decrease to communicative prosody
- The magnitude of F0 increase/decrease is in proportion to the degree of adverbs as shown in Figure 1

Moreover, we showed that the above F0 control characteristics coincided with perceptual naturalness evaluation of communicative speech using MOS (Mean Opinion Score). It was turned out that communicative F0 pattern could be obtained using the mapping from subjective degree of adverb to F0 control parameters. By the perceptual experiment evaluating naturalness of synthesized speech, the effectiveness of communicative prosody generation scheme has been confirmed.

This study showed the possibility of communicative prosody control using attributes of output lexicons, *i.e.* positive / negative of adjectives and degree of adverbs. Though there seem to exist quite many factors affecting communicative prosody, some of them can be explained as a part of linguistic attributes of lexicons constituting an utterance. This possibility of lexicon-driven communicative prosody generation has been generalized by our series of works using perceptual impressions to characterize communicative prosody [3]-[11].

2.3. Communicative prosody description using perceptual impressions

To discriminate and quantify the differences of prosody observed in the real world communication, we do not have any description framework. Though existing prosody descriptions such as ToBI can be served as a description scheme of prosody shape by itself, we do not have a description system to distinguish them as speech with some specific communicative role. We need a description scheme of communicative functions which cause these distinctions and directly relate to information manifested as prosodic differences. If we can successfully define the description of communicative prosody, we can analyze the mapping from the description to the prosody manifested in communicative speech. I have been suffering this communicative prosody specification problem for more than two decades till I found one way out through the analysis of "uhm".

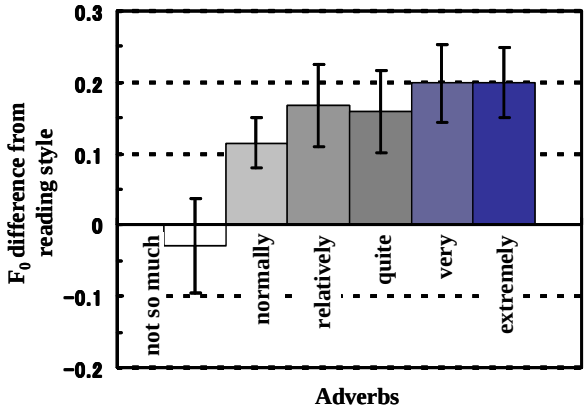

Figure 1 *The increase of F_0 average difference between reading-style and communicative one in proportion to the increase of degree of adverbs when positive adjectives follow*

Though a single utterance "uhm" does not have any particular lexical meaning by itself, its intonation can convey various kinds of communicative information. In order to treat the information conveyed by its communicative prosody, we proposed to employ perceptual impressions. Using "uhm" as a target, we can directly associate its prosodic characteristics manifested in F0 and duration with its perceptual impressions without being bothered by intrinsic lexical properties and linguistic structures [3].

Based on our observations of "uhm", twelve single word utterances that were controlled by three types of average F0 height (high, mid, low) and four types of F0 dynamics (rise, flat, fall rise&fall) were prepared as speech stimuli. After preliminary listening tests, we decided to use twenty-six word expressions for the description of perceptual impressions. These impression words can be classified into the following three groups, *doubtful-confident* (doubt, ambivalence, understanding, approve), *unacceptable–allowable* (deny, objection, agreement) and *negative- positive* (dark, weakly, not interested, bad mood, heavy, bothering, audacious, anger, annoying, cheerful, delight, gentle, good mood, excited, happy, light, interested, bright) We asked five subjects to evaluate twelve “uhm” utterances in terms of twenty-six impression word with eight-level scaling, 0(not at all)- 7(very much).

By applying Multi-Dimensional Scaling (MDS), the above perceptual impression space with twenty six dimensions could be reduced to three dimensions. To interpret the meaning of obtained three axes obtained by MDS, the average scores corresponding to each impression word were projected onto the three-dimensional spaces. In Figure 2, we plotted the impression words that showed high correlation between these three axes. As shown in this figure, we can approximate a set of impression words in three dimensions expressing the

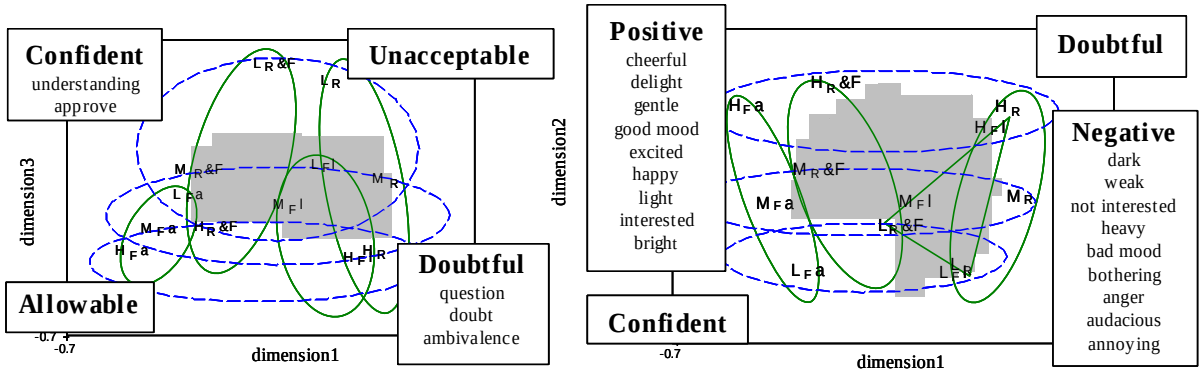

Figure 2 *Projection of test word vectors in three dimensional perceptual impression space obtained by Multi-Dimensional Scaling (INDSCAL)* (Clusters of F0 height (H,M,L) and F0 dynamic patterns (R,Fl,Fa,R&F) are circled with dash line and full line respectively)

speaking attitudes of *positive–negative*, *confident–doubtful* and *allowable–unacceptable*. The axes of *confident–doubtful* and *positive–negative* can be projected on the plane spanned by the first and second dimensions. The axes of *allowable–unacceptable* and *confident–doubtful* can be interpreted in the plane spanned by the first and third dimensions. The axes of *allowable–unacceptable* and *positive–negative* are interpreted in the plane of the second and third dimensions. These results nicely coincide with our intuitive grouping of the twenty-six basic expressions given in the previous paragraph and support the possibility of treatments of perceptual impressions by a restricted number of freedoms conveyed just by F0 average height and shapes.

2.4. Communicative prosody generation using impression attributes of output lexicons

The studies in the previous section suggested further possibilities of impression attributes as input to specify communicative prosody. As "uhm" has not specific meaning by itself, if we want to generate some specific "uhm" with communicative prosody, we need to assign its impression. While, if we want to generate some word utterance directly associated to impressions observed in "uhm" analysis, it may be possible to get those impressions directly from the lexicon by itself. That is, the default impression might be obtained from the lexicon by itself. To confirm this idea of the relationship between word impressions and communicative prosody, we have observed the communicative prosody of simple phrase expressions corresponding to a three-dimensional space.

We collected communicative speech data of sixteen common Japanese phrases (*doubtful-confident*: doubt, ambivalence, understanding, approve, *unacceptable-allowable*: deny, objection, agreement, sympathy, *negative-positive*: dark, sad, not interested, heavy, bright, happy, interested, and light). We observed their prosodic characteristics in conversational speech [4]. The correlations analyses showed that word impressions directly corresponding to three dimensions in a perceptual impression space had the same prosodic characteristics of "uhm" showing the corresponding impressions. The word attributes expressing *confident-doubtful; allowable-unacceptable* could be dependent on the difference of F0 dynamic patterns and duration, while those of *positive-negative* were highly related with the F0 height. These results showed the usefulness of the word impressions for communicative prosody generation.

These correlations between word impressions and communicative prosody characteristics have been confirmed not only single word utterances but also a phase utterances consisting of multiple words with different impressions [11]. From these observations, we can think of a communicative prosody generation scheme as shown in Figure 3. As shown in this figure, input lexicons are used not only for the calculation of conventional prosody such as phrasing and phrase accents but also for the calculation of communicative contributions. For the F0 generation, we employed the command-response model [12] where conventional prosody control and the communicative one could be added in its control parameter domain [6]. Perceptual naturalness evaluation experiments on synthesized speech with communicative prosody have shown the usefulness of this communicative prosody control [5].

2.5. Possibilities of language universal prosody control using impression-prosody correlation

From the impressions employed for communicative prosody control, we can easily guess the applicability of the proposed lexicon-driven communicative prosody control to other languages. To confirm the applicability of the proposed impression- prosody mapping to other languages, we carried out experiments on the communicative prosody generation for Mandarin and English phrases using the proposed generation scheme [7], [10]. By using corresponding utterances of "uhm" uttered by a Japanese speaker, we extracted communicative prosody component using the command response model. For both Chinese and English words directly showing the impressions corresponding to six axes of three dimensions were selected.

By analyzing read speech of these Chinese and English words using the command response model fitting, their word intrinsic prosody parameters were obtained. By modifying each parameter using corresponding communicative prosody parameters extracted from of "uhm" uttered by a Japanese speaker, the communicative speech samples were synthesized using STRAIGHT synthesis [13]. Perceptual evaluation experiments of synthesized speech showed remarkable naturalness increase of the synthesized communicative

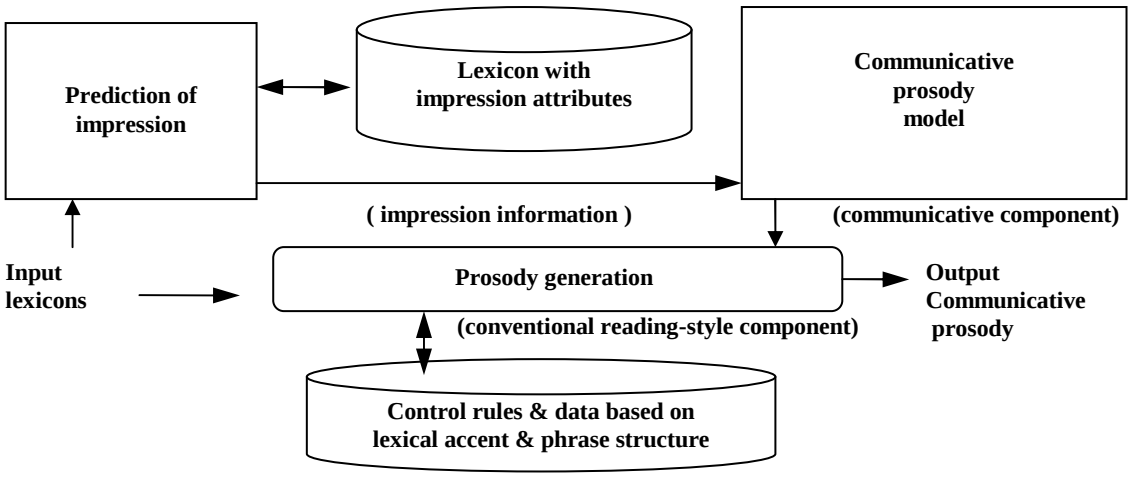

Figure 3 *Communicative prosody generation using impression prediction by input lexicons*

prosody samples. Though we could have confirmed the applicability to representative words directly associated impressions derived from MDS analyses only, we could speculate the generality of this communicative prosody control.

2.6. Further steps towards dialogue prosody control

As shown in above sections, we could have confirmed the possibility of communicative prosody control based on impression-prosody correlations of constituent lexicons. However, it is also true that the communicative prosody cannot be simply obtained from constituent lexicons. For example, compare the following two utterances in conversation.

- Their rooms were so dirty!
- Are their rooms dirty or clean?

Though the same word "dirty" is employed in these two utterances, their communicative prosody is not identical. The first example shows the speaker's negative opinion to the room using the lexicon "dirty". On the other hand, "dirty" in the second one has one of the possibilities of the status of the room and does not indicate speaker's opinion. As shown the differences in these examples, we have to pay attention to the communicative function of the utterance by itself not only sticking to the impression attributes of constituent words.

In the field of spoken dialogue, statistical models have been proposed to identify the dialogue act and its structure [14]-[16]. Free from taxonomic classification and identification problems of dialogue acts, these statistical dialogue models have a potential to provide a soft description of the utterance act and are expected to be useful for dialogue prosody control. More intensive systematic studies are expected to control dialogue prosody to properly reflect dialogue act.

3. Studies on L2 timing characteristics

In contrast to F0, we treat variations of timing characteristics due to language differences in this section. The deviations from native speaker's timing characteristics are analyzed to characterize the second language learner's speech timing control. Their perceptual characteristics are also discussed to understand human's subjective quality evaluation properly.

3.1. Statistical duration modeling

Temporal control characteristics of speech have been studied for a long time for many purposes in different speech-related fields. Segmental duration characteristics have been measured over many languages to understand language universal/specific prosodic control by many phonetic scientists. In speech technology, fine control of segmental duration has been pursued to synthesize speech with natural rhythm and tempo. Based on the characteristic analyses of segmental durations, corpus-based statistical models have been studied [17]-[28]. The following linear regression model has been widely used [18]-[22][25][26].

$$DUR = \mu(/*/) + \sum_{f} \sum_{c} Xfc\,\delta fc$$

In this equation, $\mu(/*/)$ denotes the mean duration of the current phoneme */*/, Xfc corresponds to the contribution coefficient of each category c of control factor f and δfc stands for the characteristic function of category c (i.e. δfc is 1 iff the current context corresponds to category c of f, otherwise 0) The control factors { f } correspond to current and neighboring phoneme categories. For example, the following factors were employed for English duration control, the current and four context phones; stress; phone positions in each syllable, word, and phrase; the numbers of constituent phones in each syllable, word, and phrase; syllable positions in word, and phrase; the numbers of syllables in word, and phrase; and the narrow and broad parts of speech whose contribution is confirmed through statistical analyses [25][26]. By adopting the least-square error minimization criterion, modeling coefficients Xfc representing the contributions of the control factors are obtained using training data.

To effectively reduce the control freedom in regression trees by partially imposing constraints of linear models, Constrained Tree Regression (CTR) has been proposed [24]. In the CTR model, a superset of the traditional models, a regression tree [23] is generated by controlling the tiedness of control factor parameters. By untying a shared parameter, or

splitting one of the current leaves according to finer factor differences, more efficient use can be made of a new additional parameter freedom. By controlling the tiedness of the control parameters, CTR incorporates both linear and tree regressions as special cases and interpolates between them.

3.2. Segmental duration difference as an objective evaluation measure for L2 proficiency

Computational models for duration have been developed for originally text to speech synthesis. These models can also be served for the understanding of control factors and mechanisms. In particular, it can be used as a reference of native's timing characteristics. Without human's real utterance, a duration model can give any segmental duration in arbitrary context based on based on native's characteristics. Using duration calculated by the model as a reference, the timing characteristics of the second language learners can be evaluated. [26]-[29]

Figure 4 shows the duration differences from predicted durations between English native speakers and Thai learners grouped by English-education experience in English-as-an-official-language countries [26]. As shown in this Figure, noticeable duration differences were observed by learners' grouping according to the time spent in English education. The duration differences of the close speaker open-text set showed the least difference from that of the training set (i.e., the close-speaker close-text set). This group also showed the smallest duration differences among all speaker groups. Accordingly, the results showed consistency and reasonable prediction accuracy of the model both for the training and for the open set.

3.3. Perceptual timing distortion and loudness

For the evaluation of timing differences, a series of analyses on perceptual characteristics have been carried out by Kato [30]-[33]. From his studies, we know that listeners' MOS scores of acceptability against changes in segmental duration can be accurately traced by a parabolic curve and that the absolute value of the second-order coefficient of this approximation curve, (hereafter we call this *vulnerability index*) is generally larger for vowel segments than for consonant segments.

Furthermore, it has been found that this variation in the vulnerability index is highly correlated with the intrinsic loudness of the segments as shown in Figure 5 (the right-hand scale). A non-speech study on temporal discrimination capability, on the other hand, showed that an auditory duration with large loudness is more accurately discriminated than a softer duration, if the target duration is temporally flanked by other sounds [33]. This tendency in temporal discrimination capability agrees with that of the acceptability measure found in Figure 5. All of these results suggest that the correlation observed between the vulnerability index (a sensitivity measure for acceptability) and the segment loudness can be accounted for as a reflection of the general characteristics of auditory perception. To take into account these perceptual characteristics, i.e., the dependency of duration sensitivity on segment quality, for distortion evaluation, the loudness characteristics should be added to approximate human subjective judgment for timing distortions more precisely.

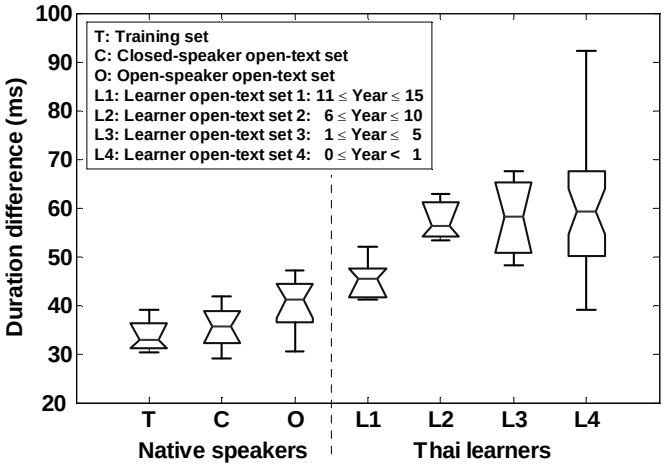

Figure 4 *Comparison of RMS duration differences from predicted durations between English natives and Thai learners*
(C: native closed speakers, O: native open speakers, L1 – L4: Thai learners grouped by education period in English-as-an-official-language countries)

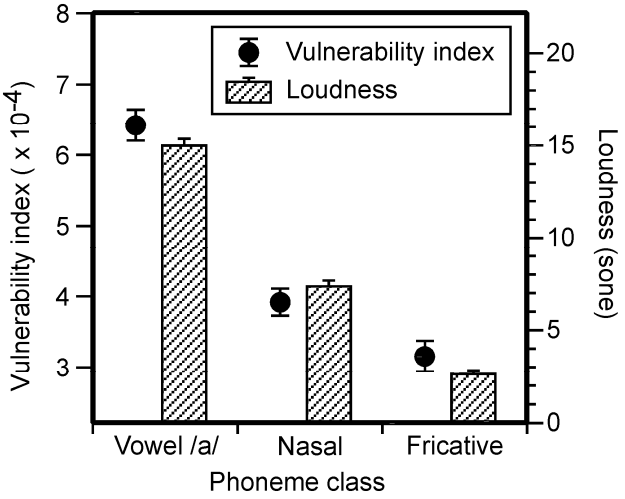

Figure 5 *The change of subjective score in relation to duration modification*
(The temporal vulnerability (the second-order coefficient of a parabolic fitting to acceptability rating scores with change in the segmental duration; dots, left-hand scale) and the loudness (bars, right-hand scale) of a speech segment as a function of phoneme class. The bars show the standard distortions. A larger vulnerability index implies a lower perceptual acceptability for a given change in the segmental duration.)

3.4. Perceptual timing distortion measure for L2 proficiency

By accumulating the above knowledge on loudness correlations to the naturalness evaluation of timing distortions, Kato proposed the time-loudness marker model for the perceptual naturalness evaluation[34][35]. Motivated by this evaluation, perceptually weighted measure for L2 speech was proposed [29]. Figure 6 shows the perceptual weighting parameters for the [k] segment in the utterance of "Thank you". The duration difference Δ k between learner's speech and the

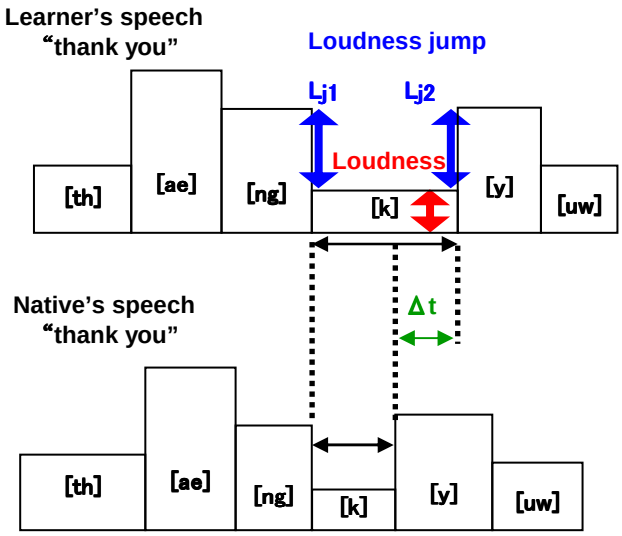

Figure 6 *Duration difference weighting using the loudness of the current segment and the loudness jumps from adjacent segments*
(Duration difference Δt is weighted by w (Lc, Lj1, Lj2) using loudness of the current segment Lc and Loudness jump from adjacent segments Lj1 & Lj2)

native's of [k] segment is weighted by the loudness Lc of [k] segment by itself and the loudness jumps Lji, Lj2 between two adjacent phones [ng] and [y].

Applying this new loudness weighted difference measure to Japanese learner's English speech, correlation score 0.54 between the raw duration differences and subjective MOS scores increased to 0.72 [29]. Moreover, this measure was used to correlate the learning difficulty of distinguishing singleton/geminate consonant clusters of Japanese for Korean learners [36]. In slow rate speech, the correlation score 0.79 between consonant length and the distinguishing error rate increased to 0.91 by introducing loudness weighting. These remarkable correlation increases indicate the usefulness of our knowledge integration for the understanding of subjective evaluation of timing characteristics.

4. AESOP research consortium

To carry out the research introduced in this paper, we have been using a lot of different kinds of speech and language corpora. We could have got none of these research results without them. To obtain full annotated rich data, it takes huge human efforts and time for content design, data collection, various types of transcriptions and preparation for associated information such as speaker's background and natives' subjective evaluation scores for L2 data. We will definitely need more data to expand our research area to neighboring related field of education or other basic science.

To facilitate our research and enhance collaborations in different countries and multi-disciplinary research fields, in particular L2 related fields, we set up the consortium of AESOP (Asian English Speech cOrpus Project) where researchers in Asian countries have started to work together in 2008 fall at Waseda in Japan. The first target of AESOP is to build up a common English speech corpus which represents the varieties of English spoken in Asia. We are forming an international consortium of linguists, psychologists, speech scientists, technologists and educators from Asian countries at the beginning. We are going to collect and compare English speech corpora first from the Asian countries using a consistent set of core materials in order to derive a set of phonetic properties common to all varieties of Asian English [37] and provide a research platform that we can share. We sincerely hope that we will make this research consortium as a trans-disciplinary group where researchers in different fields can communicate each other by respecting each discipline and effectively share valuable human wisdom and knowledge.

5. Conclusions

In this paper, I have introduced prosody variation modeling for communicative prosody characterization and the objective evaluation of L2 timing control characteristics. For communicative prosody modeling, a possibility of lexicon driven control and further needs of dialogue-act modeling are discussed. For L2 timing control characteristics, possibilities of objective evaluation of L2 learner's proficiency are shown using duration modeling and perceptual characteristics. Through these modeling processes, I would like to propose a rethinking of conventional notions such as "para-linguistics" and integration of our knowledge in trans-disciplinary manner to teat undiscovered interesting prosodic phenomena in scientific way. As exemplified by the usefulness of statistical duration modeling and loudness correlations on perceptual timing characteristics tell us the importance of collaboration. They are not merely useful in their own research field such as information processing or perceptual studies but also many neighboring other research areas. I am quite sure that the integration of all knowledge that we have in different fields will bring us many more benefits in every related fields.

6. Acknowledgements

This work was supported in part by Grant-in-Aid for Scientific Research B, No.20300069 and No. 23320091 of JSPS. The author would like to express sincere thanks to many collaborators. In particular, Hiroaki Kato, Minoru Tsuzaki, Takumi Yamashita, Makiko Muto, Yoko Greenberg, Ming Zhu, Ke Li, Chatchawarn Hansakunbuntheung, Shizuka Nakamura, Mee Sonu and Hajime Tsubaki for their original contributions and assistances. As shown in the references, this paper consists of their original works and gives a unified view underlying these works.

7. References

[1] Sapir E., "Language: An introduction to the study of speech" Harcourt, Brace and Company 1921.

[2] Sagisaka Y., Yamashita T. and Kokenawa Y. , "*Generation and perception of F0 markedness for communicative speech synthesis*", Speech Communication, Vol.46, No.3-4, pp.376-384 2005.

[3] Kokenawa Y., Tsuzaki M., Kato H. and Sagisaka Y., *"F0 control characterization by perceptual impressions on speaking attitude using Multiple Dimensional Scaling analysis"* Proc. IEEE ICASSP 2005 pp.273-275.2005.

[4] Kokenawa, Y., Tsuzaki, M., Kato, K., and Sagisaka, Y. "*Communicative speech synthesis using constituent word attributes"*, Proc. INTERSPEECH, 517-520 2005.

[5] Greenberg, Y., Tsuzaki, M., Kato, K., and Sagisaka, Y., "*A trial of communicative prosody generation based on control characteristic of one word utterance observed in real conversational speech*", Proc. Speech Prosody 2006, pp.37-40, 2006

[6] Li K., Greenberg, Y., Campbell N. and Sagisaka Y.," *On the analysis of F0 control characteristics of nonverbal utterances and its application to communicative prosody generation*" in NATO Security through Science Series E: Human and Societal Dynamics Vol.8 The Fundamentals of Verbal and Non-verbal Communication and the Biometric Issue pp.179-183 IOS Press 2007

[7] Li K., Greenberg, Y. and Sagisaka Y., "*Inter-language prosodic style modification experiment using word impression vector for communicative speech generation*", Proc. Interspeech 2007 pp.1294 -1297 2007

[8] Zhu M., Li K., Greenberg and Sagisaka Y., "*Automatic extraction of paralinguistic information from communicative speech*" Proc. the 7th Symposium on Natural Language Processing 2007 pp.207-212 Dec.2007

[9] Greenberg Y., Shibuya N., Tsuzaki M., Kato H., and Sagisaka Y., "*Analysis on paralinguistic prosody control in perceptual impression space using multiple dimensional scaling*", Speech Communication Vol.51 No.7 pp. 585-593, 2009.

[10] Greenberg Y., Tsuzaki M., Kato H., and Sagisaka Y., "*Communicative prosody generation using language common features provided by input lexicons*", Proc. SNLP2009, pp.101-104, 2009.

[11] Greenberg Y., Tsuzaki M., Kato H., and Sagisaka Y., "*Analysis of impression-prosody mapping in communicative speech consisting of multiple lexicons with different impressions*", Proc. O-COCOSDA, 2010

[12] Fujisaki, H. and Hirose, K., "*Analysis of voice fundamental frequency contours for declarative sentences of Japanese*," J. Acoust. Soc. Japan (E), Vol.5, No.4, pp.233-242, 1984.

[13] Kawahara, H., Masuda-Katsuse, I. and Cheveign´e, A., "*Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency based F0 extraction: Possible role of a repetitive structure in sounds*," Speech Communication Vol.27, pp.187-207, 1999

[14] Young S., "*Cognitive User Interfaces: an Engineering Approach*," Pleanary Talk in .IEEE ICASSP 2009

[15] Young S., Gasic M., Keizer S., Mairesse F., Schatzmann J., Thomson B. and Yu K., "*The Hidden Information State Model: a practical framework for POMDP-based spoken dialogue management.*" Computer Speech and Language, 24(2): 150-174. 2010.

[16] Quarteroni S., Ivanov A. V., Riccardi G.," *Simultaneous dialog act segmentation and classification from human-human spoken conversations.*" Proc. ICASSP pp. 5596-5599 2011

[17] Sagisaka Y. and Tohkura Y., "*Phoneme duration control for speech synthesis by rule*" (in Japanese) Trans. IElCE J67-A, No.7, pp.629-636, 1984

[18] Takeda K , Sagisaka Y and Kuwabara H., "*On sentential effects in the control of segmental duration in Japanese*" JASA Vol.86 (6) pp.2081-2087, 1989

[19] Kaiki N., Takeda K and Sagisaka Y., "*Linguistic properties in the control of segmental duration for speech synthesis*" p.255-264 in "Talking Machines" edited by Bailly G. et al North-Holland, 1992

[20] Kaiki N. and Sagisaka Y., "*The control of segmental duration in speech synthesis using statistical models*" pp.391-402 in "Speech perception, production and linguistic structure" edited by Tohkura Y.et al Ohmsha IOS press, 1992

[21] Sagisaka Y., "*Prosody control for spontaneous speech synthesis*" Proc. ICPhS pp.506-509, 1991

[22] Kaiki N. and Sagisaka Y., "*Pause characteristics and local phrase-dependency structure in Japanese*" Proc.ICSLP92 pp.357-360, 1992

[23] Riley M.D.: "*Tree-based modeling of segmental durations*" p.265-274 in "Talking Machines" edited by Bailly G. et al North-Holland, 1992

[24] Iwahashi N. and Sagisaka Y.: "*Statistical modeling of speech segment duration by constrained tree regression*" Trans. IEICE Vol.E83-D, pp.1550-1559, 2000

[25] Hansakunbuntheung C., Kato H. and Sagisaka Y., "*Syllable-based Thai Duration Model Using Multi-level Linear Regression and Syllable Accommodation*" Proc. the 6th ISCA Speech Synthesis Workshop pp.356-361 2007

[26] Hansakunbuntheung C., Kato H. and Sagisaka Y., "*Model-based automatic evaluation of second-language learner's English segmental duration characteristics*" J. Acoust. Sc. Tech. Vol. 31 No. 4 pp.267-277, 2010

[27] Muto, M., Sagisaka Y., Naito T., Maeki D., Kondo A., Shirai K., "*Corpus-based modeling of naturalness estimation in timing control for non-native speech*" Proc. EUROSPEECH pp.498-501, 2003

[28] Nakamura S., Tsubaki H., Kondo Y., Nakano M. and Sagisaka Y.,"*Tempo-normalized measurement and test set dependency in objective evaluation of English learners' timing characteristics*" Proc. 16th ICPhS pp.1733-1736, 2007

[29] Nakamura S., Matsuda S., Kato H., Tsuzaki M. and Sagisaka Y., "*Objective evaluation of English learners' timing control based on a measure reflecting perceptual characteristics*" Proc. IEEE ICASSP pp.4837-4840, 2009

[30] Kato H., Tsuzaki M. and Sagisaka Y., "*Acceptability for temporal modification of consecutive segments in isolated words*" JASA. Vol. 101, pp.2311-2322, 1997

[31] Kato H., Tsuzaki M. and Sagisaka Y., "*Acceptability for temporal modification of single vowel segments in isolated words*," JASA Vol. 104, pp.540-549, 1998

[32] Kato H., Tsuzaki M. and Sagisaka Y., "E*ffects of phoneme class and duration on the acceptability of modifications in speech*" JASA. Vol. 111, pp. 387-400, 2002

[33] Kato H. and Tsuzaki M., "*Intensity effect on discrimination of auditory duration flanked by preceding and succeeding tones*" JASJ (E) Vol.15, pp.349-351, 1994

[34] Kato, H., Tsuzaki, M., and Sagisaka, Y., "*A modeling of the objective evaluation of durational rules based on auditory perceptual characteristics.*" Proc. ICPhS, 1835-1838. 1999.

[35] Kato, H., Tsuzaki, M., and Sagisaka, Y. "*A modeling of the objective evaluation for durational rules based on auditory perceptual characteristics*." J. Acoust. Soc. Jpn, Vol. 55-11, pp.752-760. 1999. (in Japanese with English abstract and figure captions.)

[36] Sonu M., Tajima K., Kato H. and Sagisaka Y.,"*Perceptual studies of Japanese geminate insertion phenomena based on timing control characteristics*" Proc. ICPhS 2011 pp.1886-1889, 2011

[37] Visceglia T., Tseng C., Kondo M., Meng M. and Sagisaka Y.,"*Phonetic Aspects of Content Design in AESOP (Asian English Speech cOrpus Project)*" Proc. O-COCOSDA pp.60-65, 2009