Editorial. Special issue on Quantitative prosody modeling for natural speech descriptionand generation

Keikichi Hirose, Daniel J. Hirst, Yoshinori Sagisaka

Keikichi Hirose, Daniel Hirst and Yoshinori Sagisaka. 2005. Editorial (special issue: *Quantitative Prosody Modeling for Natural Speech Description and Generation*. *Speech Communication* 46 (3-4), 217–219

doi:10.1016/j.specom.2005.05.006

In human communication through spoken language, segmental features serve a major role in the transmission of linguistic information of word meanings and, therefore, utterance contents. In contrast, prosodic features underlie higher-level information at the syntactic and discourse levels, and dominate the expression of attitudes, emotions, and affective information. Reflecting the importance of prosody in human verbal communication, a number of research projects are under way at academic and industrial research organizations throughout the world. Speech prosody covers a huge multidisciplinary area involving academics, scientists, and engineers with various research backgrounds, united by an interest in human communication. Prosodic features play an important role in speech and language research, and the current developments in speech technology call for further interdisciplinary work that builds on this diverse variety of inputs. This situation increased a desire amongst speech researchers to share their knowledge in the field of speech prosody, and led to the international conference on Speech Prosody held in April 2002 in Aix-en-Provence, France. Because of its great success, many researchers working in the domain of speech prosody looked forward to its sequels, and the second conference on Speech Prosody was held in March 2004 in Nara, Japan. One hundred and seventy three papers were presented in 6 oral and 9 poster sessions. Although they covered various topics on speech prosody, modeling prosody in a quantitative way and properly controlling it in speech communication systems such as speech synthesis and recognition were one of the major concerns for the conference attendees. As exemplified by many advances of prosody control in speech communication derived from generation modeling for conventional application uses, acoustic analyses with an underlying model and their quantitative descriptions have been quite important in realizing speech communication with natural prosody. Taking these situations into account, this special issue was organized. We carefully selected papers

related to the topic of the special issue from those presented at Speech Prosody 2004, and asked the authors for extended versions. As a result, 16 papers are included in the issue on: analysis and modeling of prosody, control of prosody in speech synthesis, and use of prosody in speech recognition. The scope of the papers covers not only reading style utterances, but also emotional speech, singing voice and interaction with vision. In the first paper by Xu, prosody was viewed from two aspects: information transmitted by prosody, and acoustic manifestations of prosody. This view leads to a comprehensive model of tone and intonation, the PENTA model. These two aspects of prosody are often mixed up in other prosody modeling schemes, leading to a less clear understanding of speech prosody. Bänziger and Scherer analyzed how emotions influence prosodic features. They showed that the level and range of fundamental frequency (F0) change systematically by the degree of emotion activation, while the shape of F0 contour does not clearly differ depending on the emotion types. The samples analyzed are with ``portrayed'' emotions. House examined the extent of final F0 rise in wh-question utterances extracted from a spontaneous speech corpus in Swedish. Percentages of final rises show some differences according to gender and age. The role of final rises for transmitting friendliness and social interest in wh-questions is shown through perceptual experiments. Tseng et al. viewed the prosody of Mandarin continuous speech in a range wider than a sentence and constructed a hierarchical representation of phrase group prosody. A model for fluent speech prosody is given based on the representation. Mixdorff and Pfitzinger compared prosodic features of dialogue (spontaneous) speech and its reading version. F0 contours were parameterized based on the generation process model and analyzed. Comparison was also conducted on the speech rates. The results on accent commands mostly coincide with those on Japanese. Carlson, Hirschberg and Swerts examined if listeners could predict upcoming prosodic boundaries. Perceptual experiments were conducted for native speakers of Swedish and American English using fragments of spontaneous Swedish speech. Experiments showed the speakers of both languages could predict the boundaries well alike. Hirst gave a good view of prosody, which was similar to that by Xu; two levels of prosodic function and form need to be distinguished in prosody modeling. Their annotation systems, IF and INTSINT, are introduced together with an automatic annotation algorithm, MOMEL. Bailly and Holm explained their SFC model in detail. The model assumes an F0 contour as a sum of multiple levels of F0 movements, which are related to (meta-)linguistic information using neural networks. The model has a high flexibility in its framework, such as the number of levels, and so on. Van Santen et al. gave another super-positional model. An F0 contour in log scale is represented as a sum of phrase, accent and segmental perturbation curves. Different from the SFC model, their model includes phrase curve parameters, thus adding non-additive features into the model. Two versions of the modeling are presented. Sagisaka, Yamashita and Konekawa examined how F0 contours changed in conversational situations depending on the degree of markedness, which was controlled by the preceding adverb in adjective phrases. F0 analyses and perceptual experiments were conducted when the adjective had positive/negative impression. Corpus-based generation of F0 contours was realized using the generation process model. Hirose et al. developed a fully automatic scheme of corpus-

based generation of F0 contours from text under the framework of generation process model. The method includes automatic extraction of model parameters. Experiments were shown for several types of emotion with objective evaluation of synthetic speech quality. Saitou, Unoki and Akagi showed the importance of fluctuations in F0 for singing voice quality. Based on the results of analyses and perceptual experiments, they developed an F0 control model, which could properly handle F0 fluctuations. Hasegawa-Johnson et al. presented schemes to improve automatic labeling of prosody and to reduce word error rate in automatic speech recognition using prosodic features. A dynamic Bayesian network model, with hidden variables of words, prosodic tags and prosody-dependent allophones, was shown with experimental results. Zhang, Nakamura and Hirose achieved high performance in HMM-based tone recognition of Standard Chinese. They succeeded to suppress tone coarticulation effects by viewing tone nuclei instead of whole syllable parts and by taking carry over and anticipatory effects from the adjacent syllables into account. Shriberg et al. developed a new approach to modeling idiosyncratic prosodic features for speaker recognition. They quantized syllable F0, duration and power features, and modeled their N-gram counts using support vector machines. Its effect on speaker recognition was shown when it was combined with a baseline speaker recognition system of cepstral Gaussian mixture model. Granström and House showed the importance of audio-visual aspects of prosody in verbal communication through their experiments using dialogue systems with talking heads. They showed that the timing of gestures played an important role in expressive speech communication. This special issue is only possible with the cooperation of authors and guest reviewers. We would like to express our sincere appreciation to them. The issue will offer SPECOM readers a good prospect on how prosody should be handled in future spoken language technology.

Authors

- *Keikichi Hirose* Department of Information and Communication Engineering Graduate School of Information Science and Technology The University of Tokyo 7-3-1 Hongo, Bunkyo-ku Tokyo 113-0033, Japan
  E-mail address: hirose@gavo.t.u-tokyo.ac.jp
- *Daniel Hirst* CNRS Laboratoire Parole et Langage Universite´ de Provence 29 avenue Schuman 13621 Aix-en-Provence Cedex 1 France
  E-mail address: daniel.hirst@lpl.univ-aix.fr
- *Yoshinori Sagisaka* Graduate School of Global Information and Telecommunication Studies Waseda University 1-3-10 Nishi Waseda Shinjuku-ku 169-0051 Japan
  E-mail addresses: sagisaka@giti.waseda.ac.jp yoshinori:sagisaka@atr:jp