Selected Innovations, Trends, Issues, and Challenges in Prosody

. . . with some starting points for further exploration

multistream temporal configurations . . . Byrd and Saltzman (1998); Kochanski (2006); Ogden (2010, 2012); Day-O'Connell (2013); Niebuhr (2015); Torreira and Valtersson (2015); Ward, Carlson, and Fuentes (2018); Ward (2019); Niebuhr and Neitsch (2019); Niebuhr (2019)

early fusion . . .

gradient, not categorical . . . Kurumada, Brown, and Tannenhaus (2012); Kurumada and Clark (2017); Grice et al. (2017); Ward and Jodoin (2019)

the power of prosody . . .

interpersonally-attuned prosody . . .

context-aware prosody . . .

engineered feature sets and salads . . . Eyben, Wöllmer, and Schuller (2010); Huang, Chen, and Harper (2006); Ferrer, Scheffer, and Shriberg (2010); Eyben et al. (2016); Ward (2017)

feature parsimony . . . Skantze (2017); Ward et al. (2018)

jointly-enacted prosodic patterns . . . Lerner (2002); Corps, Gambi, and Pickering (2018); Ward and Abu (2016); Ward (2019)

superpositional modeling . . . Kochanski and Shih (2003); Fujisaki (2004); van Santen, Mishra, and Klabbers (2004); Gubian, Boves, and Cangemi (2011); Xu (2011); Liu and Xu (2016); Gerazov and Bailly (2018); Ward (2019)

the almost-independence of prosody . . . Bulyko and Ostendorf (2001); Calhoun and Schweitzer (2012); Torreira and Grice (2018); Vigario, Cruz, and Frota (2019)

sprawling prosodic meanings . . . Hedberg et al. (2010); Ward (2019)

pitch perceptions beyond simple F_0-based measures . . . Terken (1993); Kochanski (2010); Mertens (2004); Slaney, Shriberg, and Huang (2013); Barnes et al. (2012); Bruggeman et al. (2017)

individual variation . . . Pierrehumbert and Steele (1989); Weber et al. (2002); Shriberg et al. (2005); Dediu and Ladd (2007); Wong et al. (2008); Yoon (2010); Barnes et al. (2012); Niebuhr et al. (2011); Bruggeman et al. (2017); Henriksen (2013); Jun and Bishop (2015); Cangemi, Krüger, and Grice (2015); Roy, Cole, and Mahrt (2017); Boll-Avetisyan, Bhatara, and Höhle (2017); Roessig, Mucke, and Grice (2019); Kim (2019); Buxó-Lugo and Kurumada (2019)

normalization . . . Sönmez et al. (1997); Shriberg et al. (2000); Ostendorf, Shafran, and Bates (2003); Mariooryad and Busso (2014); Ward (2019)

wide diversity of functions served by prosody . . . Wells (2006); Cole (2015)

sequence-to-sequence modeling . . . Zen, Tokuda, and Black (2009); Skerry-Ryan et al. (2018); Lee and Kim (2019)

unsupervised methods . . . Greenberg et al. (2009); Gubian, Cangemi, and Boves (2010); Janssoone et al. (2016); Madaio et al. (2017); Obin and Beliao (2018); Ward (2019); Wang et al. (2018)

diverse applications and uses . . . including speech recognition, speaker identification, intelligible speech synthesis, expressive speech synthesis, communicatively effective speech synthesis, information retrieval, summarization, assessment of language proficiency, infering speaker states such as tireness and anger, diagnosis of clinical conditions including communicative disorders, dialog systems including turn-taking and infering user mental states, stances and intentions, assessment of language proficiency, training people to speak or communicate better, processing low resource languages, etc. . . . Toyama, Saito, and Minematsu (2017) . . .

scientific questions . . . mental representation of prosodic knowledge, acquisition of prosodic knowledge and skills, cognitive processes in prosody comprehension and in production, neural pathways, specificity or relation to other language and multimodal skills, universals, etc.

challenges for speech technology and sister fields . . . Mennen (2015); Ward and DeVault (2016); Niebuhr and Ward (2018); Rosenberg (2018)

References

Barnes, J.; Veilleux, N.; Brugos, A.; and Shattuck-Hufnagel, S. 2012. Tonal center of gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology* 3:337–382.

Boll-Avetisyan, N.; Bhatara, A.; and Höhle, B. 2017. Effects of musicality on the perception of rhythmic structure in speech. *Laboratory Phonology* 8(1).

Bruggeman, A.; Cangemi, F.; Wehrle, S.; El Zarka, D.; and Grice, M. 2017. Unifying speaker variability with the tonal centre of gravity. In Belz, M., and Mooshammer, C., eds., *Phonetics and Phonology in German-speaking Countries*, 21–24.

Bulyko, I., and Ostendorf, M. 2001. Joint prosody prediction and unit selection for concatenative speech synthesis. In *IEEE ICASSP*, volume 2, 781–784.

Buxó-Lugo, A., and Kurumada, C. 2019. Encoding and decoding of meaning through structured variability in intonational speech prosody. PsyArXiv.com.

Byrd, D., and Saltzman, E. 1998. Intragestural dynamics of multiple prosodic boundaries. *Journal of Phonetics* 26:173–199.

Calhoun, S., and Schweitzer, A. 2012. Can intonation contours be lexicalised? implications for discourse meanings. In Elordieta, G., and Prieto, P., eds., *Prosody and Meaning*. De Gruyter Mouton. 271–327.

Cangemi, F.; Krüger, M.; and Grice, M. 2015. Listener-specific perception of speaker-specific production in intonation. In Fuchs, S.; Pape, D.; Petrone, C.; and Perrier, P., eds., *Individual differences in speech production and perception*. Frankfurt: Peter Lang. 123–145.

Cole, J. 2015. Prosody in context: a review. *Language, Cognition and Neuroscience* 30:1–31.

Corps, R. E.; Gambi, C.; and Pickering, M. J. 2018. Coordinating utterances during turn-taking: The role of prediction, response preparation, and articulation. *Discourse Processes* 55:230–240.

Day-O'Connell, J. 2013. Speech, song, and the minor third: An acoustic study of the stylized interjection. *Music Perception* 30:441–462.

Dediu, D., and Ladd, D. R. 2007. Linguistic tone is related to the population frequency of the adaptive haplogroups of two brain size genes, ASPM and Microcephalin. *Proceedings of the National Academy of Sciences* 104:10944–10949.

Eyben, F.; Scherer, K. R.; Schuller, B. W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L. Y.; Epps, J.; Laukka, P.; Narayanan, S. S.; et al. 2016. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing* 7:190–202.

Eyben, F.; Wöllmer, M.; and Schuller, B. 2010. OpenSmile: the Munich versatile and fast open-source audio feature extractor. In *Proceedings, International Conference on Multimedia*, 1459–1462.

Ferrer, L.; Scheffer, N.; and Shriberg, E. 2010. A comparison of approaches for modeling prosodic features in speaker recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 4414–4417.

Fujisaki, H. 2004. Information, prosody, and modeling – with emphasis on tonal features of speech. In *Speech Prosody*.

Gerazov, B., and Bailly, G. 2018. Pysfc - a system for prosody analysis based on the superposition of functional contours prosody model. In *Proc. 9th International Conference on Speech Prosody*, 774–778.

Greenberg, Y.; Shibuya, N.; Tsuzaki, M.; Kato, H.; and Sagisaka, Y. 2009. Analysis on paralinguistic prosody control in perceptual impression space using multiple dimensional scaling. *Speech Communication* 51:585–593.

Grice, M.; Ritter, S.; Niemann, H.; and Roettger, T. B. 2017. Integrating the discreteness and continuity of intonational categories. *J. of Phonetics* 64:90–107.

Gubian, M.; Boves, L.; and Cangemi, F. 2011. Joint analysis of F0 and speech rate with functional data analysis. In *ICASSP*, 4972–4975.

Gubian, M.; Cangemi, F.; and Boves, L. 2010. Automatic and data driven pitch contour manipulation with functional data analysis. In *Speech Prosody*.

Hedberg, N.; Sosa, J. M.; Gorgulu, E.; and Mameni, M. 2010. The prosody and meaning of wh-questions in american english. In *Speech Prosody*.

Henriksen, N. 2013. Style, prosodic variation, and the social meaning of intonation. *Journal of the International Phonetic Association* 43(2):153–193.

Huang, Z.; Chen, L.; and Harper, M. P. 2006. An open source prosodic feature extraction tool. In *LREC*, 2116–2121.

Janssoone, T.; Clavel, C.; Bailly, K.; and Richard, G. 2016. Using temporal association rules for the synthesis of embodied conversational agents with a specific stance. In *International Conference on Intelli-*

gent Virtual Agents. Springer. 175–189.

Jun, S.-A., and Bishop, J. 2015. Priming implicit prosody: Prosodic boundaries and individual differences. *Language and Speech* 58(4):459–473.

Kim, J. 2019. Individual differences in the production of prosodic boundaries in American English. In *International Congress of the Phonetic Sciences*.

Kochanski, G., and Shih, C. 2003. Prosody modeling with soft templates. *Speech Communication* 39:311–352.

Kochanski, G. 2006. Prosody beyond fundamental frequency. In Sudhoff, S.; Lenertova, D.; Meyer, R.; Pappert, S.; Augurky, P.; Mleinek, I.; Richter, N.; and Schliesser, J., eds., *Methods in Empirical Prosody Research*. Walter de Gruyter Berlin. 89–121.

Kochanski, G. 2010. Prosodic peak estimation under segmental perturbations. *Journal of the Acoustical Society of America* 127:862–873.

Kurumada, C., and Clark, E. V. 2017. Pragmatic inferences in context: Learning to interpret contrastive prosody. *Journal of Child Language* 44:850–880.

Kurumada, C.; Brown, M.; and Tannenhaus, M. K. 2012. Pragmatic interpretation of contrastive prosody: It looks like speech adaptation. In *Cognitive Science Conference*.

Lee, Y., and Kim, T. 2019. Robust and fine-grained prosody control of end-to-end speech synthesis. In *IEEE ICASSP*, 5911–5915.

Lerner, G. H. 2002. Turn-sharing: The choral coproduction of talk in interaction. In Ford, C. E.; Fox, B. A.; and Thompson, S. A., eds., *The Language of Turn and Sequence*. Oxford University Press. 225–256.

Liu, X., and Xu, Y. 2016. Pitch perception of focus and surprise in Mandarin Chinese: Evidence for parallel encoding via additive division of pitch range. In *Tonal Aspects of Languages*, 129–132.

Madaio, M. A.; Lasko, R.; Cassell, J.; and Ogan, A. 2017. Using temporal association rule mining to predict dyadic rapport in peer tutoring. In *Educational Data Mining*.

Mariooryad, S., and Busso, C. 2014. Compensating for speaker or lexical variabilities in speech for emotion recognition. *Speech Communication* 57:1–12.

Mennen, I. 2015. Beyond segments: Towards a L2 intonation learning theory. In Delais-Roussairie, E.; Avanzi, M.; and Herment, S., eds., *Prosody and Language in Contact*. Springer. 171–188.

Mertens, P. 2004. The Prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In *Speech Prosody 2004*.

Niebuhr, O., and Neitsch, J. 2019. Questions as prosodic configurations: How prosody and context shape the mutltiparametric acoustic nature of rhetorical questions in German. In *International Congress of the Phonetic Sciences*.

Niebuhr, O., and Ward, N. G. 2018. Challenges in the modeling of pragmatics-related prosody: Introduction to the special issue. *Journal of the International Phonetics Association* 48:1–8.

Niebuhr, O.; D'Imperio, M.; Fivela, B. G.; and Cangemi, F. 2011. Are there shapers and aligners? Individual differences in signalling pitch accent category. In *International Congress of Phonetic Sciences*, 120–123.

Niebuhr, O. 2015. Stepped intonation contours: A new field of complexity. In Skarnitzl, R., and Niebuhr, O., eds., *Tackling the Complexity of Speech*. Charles University Press. 39–74.

Niebuhr, O. 2019. Pitch accents as multiparametric configurations of prosodic features–evidence from pitch-accent specific micro-rhythms in german. In Nyvad, A. M., ed., *A Sound Approach to Language Matters: In Honor of Ocke-Schwen Bohn*. Aarhus University. 321–351.

Obin, N., and Beliao, J. 2018. Sparse coding of pitch contours with deep auto-encoders. In *Spech Prosody*.

Ogden, R. 2010. Prosodic constructions in making complaints. In Barth-Weingarten, D.; Reber, E.; and Selting, M., eds., *Prosody in Interaction*. Benjamins. 81–103.

Ogden, R. 2012. Prosodies in conversation. In Niebuhr, O., ed., *Understanding Prosody: The role of context, function, and communication*. De Gruyter. 201–217.

Ostendorf, M.; Shafran, I.; and Bates, R. 2003. Prosody models for conversational speech recognition. In *Proc. of the 2nd Plenary Meeting and Symposium on Prosody and Speech Processing*, 147–154.

Pierrehumbert, J. B., and Steele, S. A. 1989. Categories of tonal alignment in English. *Phonetica* 46(4):181–196.

Roessig, S.; Mucke, D.; and Grice, M. 2019. The dynamics of intonation: Categorical and continuous variation in an attractor-based model. *PloS one* 14(5):e0216859.

Rosenberg, A. 2018. Speech, prosody, and machines: Nine challenges for prosody research. *Proc. of Speech Prosody* 784–793.

Roy, J.; Cole, J.; and Mahrt, T. 2017. Individual differences and patterns of convergence in prosody perception. *Laboratory Phonology* 8(1):1–36.

Shriberg, E.; Stolcke, A.; Hakkani-Tur, D.; and Tur, G. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication* 32:127–154.

Shriberg, E.; Ferrer, L.; Kajarekar, S.; Venkataraman, A.; and Stolcke, A. 2005. Modeling prosodic feature sequences for speaker recognition. *Speech Communication* 46:455–472.

Skantze, G. 2017. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In *Sigdial*.

Skerry-Ryan, R.; Battenberg, E.; Xiao, Y.; Wang, Y.; Stanton, D.; Shor, J.; Weiss, R. J.; Clark, R.; and Saurous, R. A. 2018. Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. In *Machine Learning Conference*.

Slaney, M.; Shriberg, E.; and Huang, J.-T. 2013. Pitch-gesture modeling using subband autocorrelation change detection. In *Interspeech*, 1911–1915.

Sönmez, M. K.; Heck, L.; Weintraub, M.; and Shriberg, E. 1997. A lognormal tied mixture model of pitch for prosody based speaker recognition. In *Fifth European Conference on Speech Communication and Technology*.

Terken, J. 1993. Issues in the perception of prosody. In *ESCA Workshop on Prosody*, 228–233.

Torreira, F., and Grice, M. 2018. Melodic constructions in Spanish: Metrical structure determines the association properties of intonational tones. *Journal of the International Phonetics Association* 48:9–32.

Torreira, F., and Valtersson, E. 2015. Phonetic and visual cues to questionhood in French conversation. *Phonetica* 72(1):20–42.

Toyama, S.; Saito, D.; and Minematsu, N. 2017. Use of global and acoustic features associated with contextual factors to adapt language models for spontaneous speech recognition. In *Interspeech*, 543–547.

van Santen, J. P.; Mishra, T.; and Klabbers, E. 2004. Estimating phrase curves in the general superpositional intonation model. In *Fifth ISCA Workshop on Speech Synthesis*, 61–66.

Vigario, M.; Cruz, M.; and Frota, S. 2019. Why tune or text? the role of language phonological profile in the choice of strateigeis for tune-text adjustment. In *International Congress of the Phonetic Sciences*.

Wang, Y.; Stanton, D.; Zhang, Y.; Skerry-Ryan, R.; Battenberg, E.; Shor, J.; Xiao, Y.; Ren, F.; Jia, Y.; and Saurous, R. A. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *International Conference on Machine Learning*.

Ward, N. G., and Abu, S. 2016. Action-coordinating prosody. In *Speech Prosody*.

Ward, N. G., and DeVault, D. 2016. Challenges in building highly interactive dialog systems. *AI Magazine* 37(4):7–18.

Ward, N. G., and Jodoin, J. A. 2019. A prosodic configuration that conveys positive assessment in American English. In *International Congress of the Phonetic Sciences*.

Ward, N. G.; Aguirre, D.; Cervantes, G.; and Fuentes, O. 2018. Turn-taking predictions across languages and genres using an LSTM recurrent neural network. In *IEEE Spoken Language Technology Conference*, 831–837.

Ward, N. G.; Carlson, J. C.; and Fuentes, O. 2018. Inferring stance in news broadcasts from prosodic feature configurations. *Computer Speech and Language* 50:85–104.

Ward, N. G. 2017. Midlevel prosodic features toolkit. https://github.com/nigelgward/midlevel.

Ward, N. G. 2019. *Prosodic Pattterns in English Conversation*. Cambridge University Press.

Weber, F.; Manganaro, L.; Peskin, B.; and Shriberg, E. 2002. Using prosodic and lexical information for speaker identification. In *IEEE ICASSP*, volume 1, I–141.

Wells, J. C. 2006. *English Intonation: An Introduction*. Cambridge.

Wong, P. C. M.; Warrier, C. M.; Penhune, V. B.; Roy, A. K.; Sadehh, A.; Parrish, T. B.; and Zatorre, R. J. 2008. Volume of left Heschl's gyrus and linguistic pitch learning. *Cerebral Cortex* 18:828–836.

Xu, Y. 2011. Speech prosody: A methodological review. *Journal of Speech Sciences* 1:85–115.

Yoon, T.-J. 2010. Speaker consistency in the realization of prosodic prominence in the Boston University radio speech corpus. In *Speech Prosody, Fifth International Conference*.

Zen, H.; Tokuda, K.; and Black, A. W. 2009. Statistical parametric speech synthesis. *Speech Communication* 51:1039–1064.