Computational Prosody Starter Bibliography

Pitch Perception ... Szczepek Reed (2010) chapter 2; Terken (1993); Kochanski (2010); Mertens (2004); Slaney, Shriberg, and Huang (2013); Barnes et al. (2012); Albert, Cangemi, and Grice (2018); Bruggeman et al. (2017)

Voicing Properties ... Ogden (2017) chapter 4; Esling et al. (2019)

Basic Signal Processing ... Jurafsky and Martin (in press) section 26.2

Linguistic Prosody ... Hyman (2006); Wells (2006); Cole (2015)

The Almost-Independence of Prosody ... Bulyko and Ostendorf (2001); Calhoun and Schweitzer (2012); Torreira and Grice (2018); Vigario, Cruz, and Frota (2019)

Speech Synthesis ... Zen, Tokuda, and Black (2009); Skerry-Ryan et al. (2018); Lee and Kim (2019); Hodari et al. (2021); Karlapati et al. (2021); Tan et al. (2021)

Speech Recognition ... Ostendorf, Shafran, and Bates (2003); Shriberg and Stolcke (2004); Chen et al. (2005); Ward, Vega, and Baumann (2011); Toyama, Saito, and Minematsu (2017); Ryant et al. (2014); Rosenberg (2018)

Paralinguistic Prosody ... Schuller (2011); Shor et al. (2020)

Historical Context ... Crystal (1969); Fox (2000)

Prosodic Feature Sets ... Eyben et al. (2016) sections 3.1 and 3.2; Ward (2019) chapter 8; Kajarekar et al. (2004); Huang, Chen, and Harper (2006); Ferrer, Scheffer, and Shriberg (2010); Eyben, Wöllmer, and Schuller (2010); Degottex et al. (2014); Black et al. (2015); Eyben et al. (2016); Ward (2017); Lenain et al. (2020)

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

Feature-Light Modeling ... Skantze (2017)

Turn Taking ... Skantze (2017); Corps, Gambi, and Pickering (2018); Heldner et al. (2019); Ward and Abu (2016); Ward (2019, 2020); Skantze (2021)

Multistream Temporal Configurations ... Byrd and Saltzman (1998); Kochanski (2006); Ogden (2010, 2012); Day-O'Connell (2013); Niebuhr (2015); Torreira and Valtersson (2015); Ward (2019); Niebuhr and Neitsch (2019); Niebuhr (2019)

Prosodic Meanings and Functions ... Hedberg et al. (2010); Kurumada, Brown, and Tannenhaus (2012); Kurumada and Clark (2017); Grice et al. (2017); Ward (2019); Ward and Jodoin (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)

Other Applications ... Xie et al. (2009), Kang (2010), Ward, Carlson, and Fuentes (2018), Forbes-Riley and Litman (2011), Sadoughi et al. (2017)

Individual Variation ... Pierrehumbert and Steele (1989); Weber et al. (2002); Shriberg et al. (2005); Yoon (2010); Barnes et al. (2012); Niebuhr et al. (2011); Bruggeman et al. (2017); 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); Xie, Buxó-Lugo, and Kurumada (2021); Wong et al. (2020)

Cognitive and Behavioral Modeling ...

Wagner and Watson (2010); Xu (2015); Byrd and Krivokapić (2021)

Unsupervised and Self-Supervised Methods ... Greenberg et al. (2009); Gubian, Cangemi, and Boves (2010); Neiberg, Salvi, and Gustafson (2013); Janssoone et al. (2016); Madaio et al. (2017); Obin and Beliao (2018); Ward (2019); Wang et al. (2018); Shor et al. (2020)

Challenges ... Mennen (2015); Ward and DeVault (2016); Niebuhr and Ward (2018); Rosenberg (2018); Marge et al. (in press 2021)

References

Albert, A.; Cangemi, F.; and Grice, M. 2018. Using periodic energy to enrich acoustic representations of pitch in speech: A demonstration. In *Speech Prosody*, 804–808.

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.

Black, M. P.; Bone, D.; Skordilis, Z. I.; Gupta, R.; Xia, W.; Papadopoulos, P.; Chakravarthula, S. N.; Xiao, B.; Van Segbroeck, M.; Kim, J.; Georgiou, P. G.; and Narayanan, S. S. 2015. Automated evaluation of non-native English pronunciation quality: Combining knowledge-and data-driven features at multiple time scales. In *Interspeech*.

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.

Byrd, D., and Krivokapić, J. 2021. Cracking prosody in articulatory phonology. *Annual Review of Linguistics* 7:31–53.

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.

Chen, K.; Hasegawa-Johnson, M.; Cohen, A.; Borys, S.; Kim, S.-S.; Cole, J.; and Choi, J.-Y. 2005. Prosody dependent speech recognition on radio news corpus of american english. *IEEE Transactions on Audio, Speech, and Language Processing* 14:232–245.

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.

Crystal, D. 1969. *Prosodic Systems and Intonation in English*. Cambridge University Press.

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

Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S. 2014. Covarep: A collaborative voice analysis repository for speech technologies. In *IEEE ICASSP*, 960–964.

Esling, J.; Moisik, S.; Benner, A.; and Crevier-Buchman, L. 2019. *Voice Quality The Laryngeal Articulator Model*. Cambridge University Press.

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.

Forbes-Riley, K., and Litman, D. 2011. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication* 53:1115–1136.

Fox, A. 2000. *Prosodic Features and Prosodic Structure: The Phonology of Suprasegmentals*. Oxford.

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 *IEEE 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*.

Heldner, M.; Wlodarczak, M.; Beňuš, Š.; and Gravano, A. 2019. Voice quality as a turn-taking cue. In *Interspeech*, 4165–4169.

Hodari, Z.; Moinet, A.; Karlapati, S.; Lorenzo-Trueba, J.; Merritt, T.; Joly, A.; Abbas, A.; Karanasou, P.; and Drugman, T. 2021. Camp: a two-stage approach to modelling prosody in context. In *IEEE ICASSP*, 6578–6582.

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

Hyman, L. M. 2006. Word-prosodic typology. *Phonology* 23(2):225–257.

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 Intelligent 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.

Jurafsky, D., and Martin, J. H. in press. *Speech and Language Processing, 3rd Edition*. Pearson.

Kajarekar, S.; Ferrer, L.; Sönmez, K.; Zheng, J.; Shriberg, E.; and Stolcke, A. 2004. Modeling NERFs for speaker recognition. In *ODYSSEY 2004: The Speaker and Language Recognition Workshop*.

Kang, O. 2010. Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System* 38:301–315.

Karlapati, S.; Abbas, A.; Hodari, Z.; Moinet, A.; Joly, A.; Karanasou, P.; and Drugman, T. 2021. Prosodic representation learning and contextual sampling for neural text-to-speech. In *IEEE ICASSP*, 6573–6577.

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.

Lenain, R.; Weston, J.; Shivkumar, A.; and Fristed, E. 2020. Surfboard: Audio feature extraction for modern machine learning. In *Interspeech*.

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 *Proc. 10th Educational Data Mining*.

Marge, M.; Espy-Wilson, C.; Ward, N. G.; et al. in press, 2021. Spoken language interaction with robots: Research issues and recommendations. *Computer Speech and Language*.

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*.

Neiberg, D.; Salvi, G.; and Gustafson, J. 2013. Semi-supervised methods for exploring the acoustics of simple productive feedback. *Speech Communication* 55(3):451–469.

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 *Speech 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.

Ogden, R. 2017. *An Introduction to English Phonetics, 2nd Edition*. Edinburgh University Press.

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.

Ryant, N.; Slaney, M.; Liberman, M.; Shriberg, E.; and Yuan, J. 2014. Highly accurate Mandarin tone classification in the absence of pitch information. In *Proceedings of Speech Prosody*.

Sadoughi, N.; Pereira, A.; Jain, R.; Leite, L.; and Lehman, J. F. 2017. Creating prosodic synchrony for a robot co-player in a speech-controlled game for children. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*, 91–99.

Schuller, B. 2011. Voice and speech analysis in search of states and traits. In Salah, A. A., and Gevers, T., eds., *Computer Analysis of Human Behavior*. Springer. 227–253.

Shor, J.; Jansen, A.; Maor, R.; Lang, O.; Tuval, O.; Quitry, F. d. C.; Tagliasacchi, M.; Shavitt, I.; Emanuel, D.; and Haviv, Y. 2020. Towards learning a universal non-semantic representation of speech. In *Interspeech*, 140–144.

Shriberg, E., and Stolcke, A. 2004. Prosody modeling for automatic speech recognition and understanding. In *Mathematical Foundations of Speech and Language Processing, IMA Volumes in Mathematics and Its Applications, Vol. 138*, 105–114. Springer-Verlag.

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*, 220–230.

Skantze, G. 2021. Turn-taking in conversational systems and human-robot interaction: A review. *Computer Speech & Language* 67:101178.

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*.

Szczepek Reed, B. 2010. *Analysing Conversation: An introduction to prosody*. Palgrave Macmillan.

Tan, X.; Qin, T.; Soong, F.; and Liu, T.-Y. 2021. A survey on neural speech synthesis. arXiv preprint arXiv:2106.15561.

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 strategies for tune-text adjustment. In *International Congress of the Phonetic Sciences*.

Wagner, M., and Watson, D. G. 2010. Experimental and theoretical advances in prosody: A review. *Language and cognitive processes* 25(7-9):905–945.

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.; 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.; Vega, A.; and Baumann, T. 2011. Prosodic and temporal features for language modeling for dialog. *Speech Communication* 54:161–174.

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.

Ward, N. G. 2020. Ten prosodic patterns of turn-taking in Japanese conversation. In *Proc. 10th International Conference on Speech Prosody 2020*, 764–768.

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.; Kang, X.; Wong, K. H.; So, H.-C.; Choy, K. W.; and Geng, X. 2020. ASPM-lexical tone association in speakers of a tone language: Direct evidence for the genetic-biasing hypothesis of language evolution. *Science Advances* 6(22):eaba5090.

Xie, S.; Hakkani-Tür, D.; Favre, B.; and Liu, Y. 2009. Integrating prosodic features in extractive meeting summarization. In *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, 387–391.

Xie, X.; Buxó-Lugo, A.; and Kurumada, C. 2021. Encoding and decoding of meaning through structured variability in intonational speech prosody. *Cognition* 211:104619.

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

Xu, Y. 2015. Speech prosody: Theories, models and analysis. In Meireles, A. R., ed., *Courses on Speech Prosody*. Cambridge Scholars Publishing. 142–177.

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.