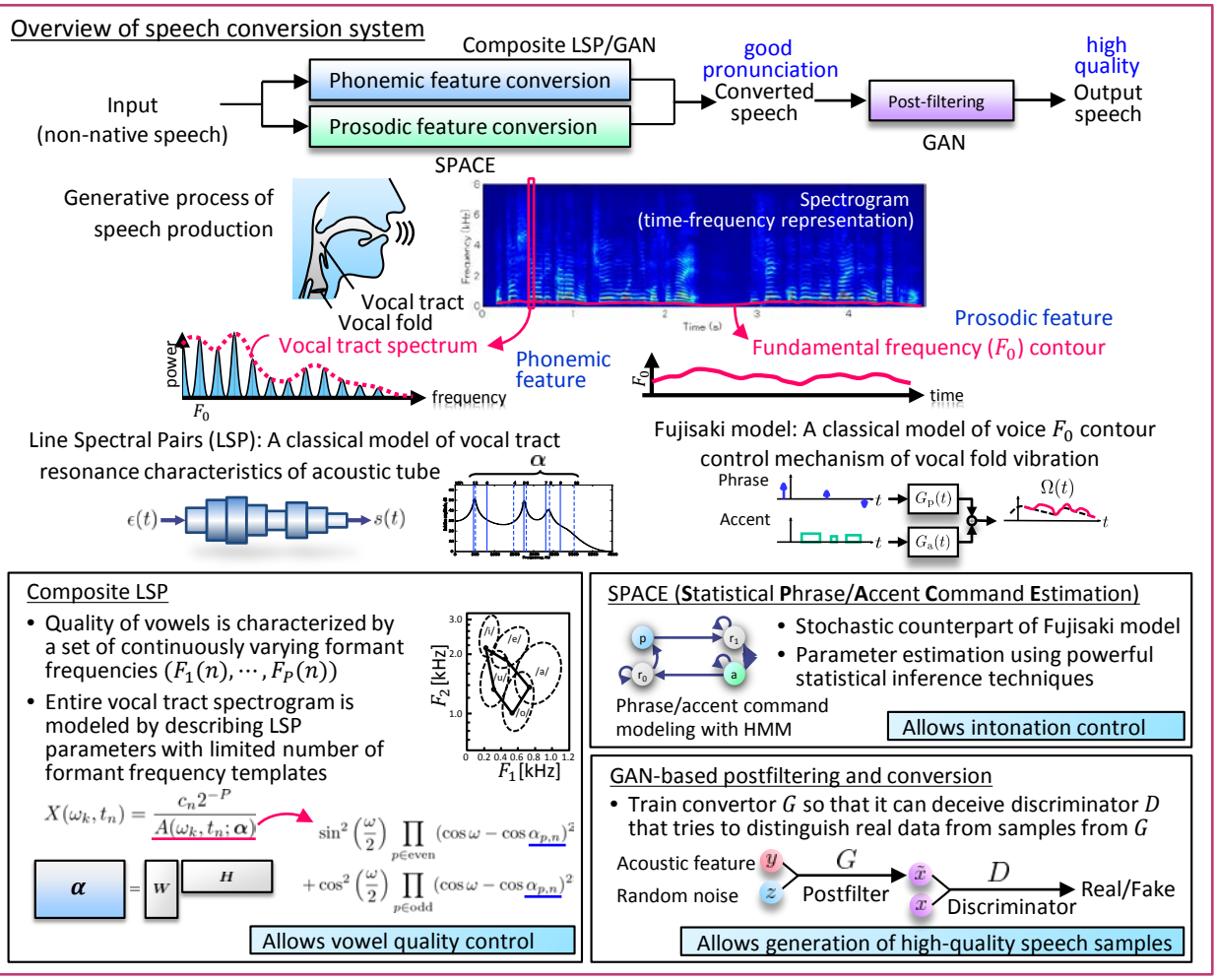
17 Converting English speech to native-like pronunciation

- Speech conversion using vocal tract model and deep generative models

Abstract

We are interested in developing a pronunciation conversion system that can convert non-native speech into intelligible native-like speech. We take a signal processing-based approach using our recently developed model, called the composite Line Spectral Pair (LSP) representation, and a deep learning-based approach using the generative adversarial network (GAN). The former approach makes it possible to convert the vowel quality of speech within physical constraints of the voice production mechanism, whereas the latter approach makes it possible to convert synthetic speech so that it becomes as indistinguishable as possible from real speech. We hope to further develop a real-time system so that it can be used to overcome many kinds of barriers to our daily communication.



Reference

[1] H. Kameoka, K. Yoshizato, T. Ishihara, K. Kadowaki, Y. Ohishi, K. Kashino, "Generative modeling of voice fundamental frequency contours," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 6, pp. 1042-1053, Jun. 2015.

[2] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *Proc. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP2017)*, pp. 4910-4914, Mar. 2017.

Contact

Hirokazu Kameoka Recognition Research Group, Media Information Laboratory
Email : kameoka.hirokazu(at)lab.ntt.co.jp