

17

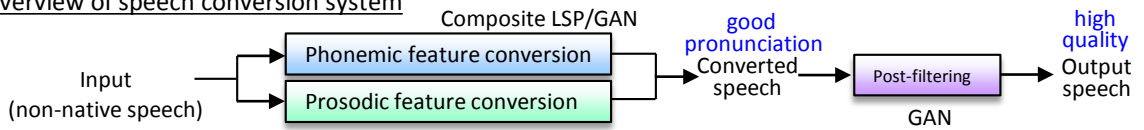
Converting English speech to native-like pronunciation

- Speech conversion using vocal tract model and deep generative models

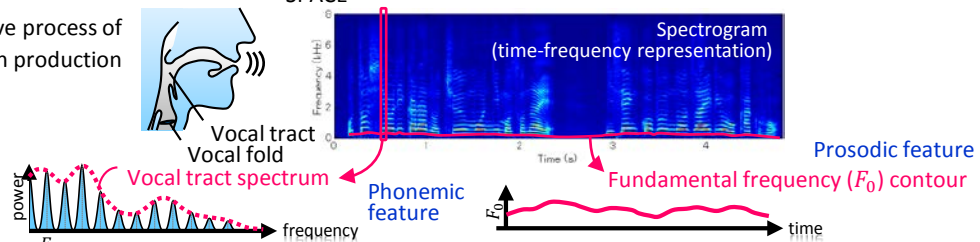
Abstract

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**.

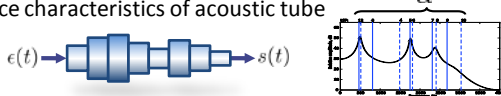
Overview of speech conversion system



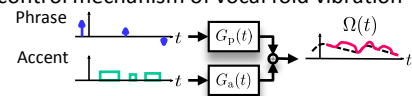
Generative process of speech production



Line Spectral Pairs (LSP): A classical model of vocal tract resonance characteristics of acoustic tube

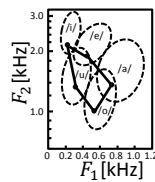


Fujisaki model: A classical model of voice F_0 contour control mechanism of vocal fold vibration



Composite LSP

- Quality of vowels is characterized by a set of continuously varying formant frequencies ($F_1(n), \dots, F_p(n)$)
- Entire vocal tract spectrogram is modeled by describing LSP parameters with limited number of formant frequency templates



$$X(\omega_k, t_n) = \frac{c_n 2^{-P}}{A(\omega_k, t_n; \alpha)} \sin^2\left(\frac{\omega}{2}\right) \prod_{p \in \text{even}} (\cos \omega - \cos \alpha_{p,n})^2 + \cos^2\left(\frac{\omega}{2}\right) \prod_{p \in \text{odd}} (\cos \omega - \cos \alpha_{p,n})^2$$



Allows vowel quality control

SPACE (Statistical Phrase/Accent Command Estimation)

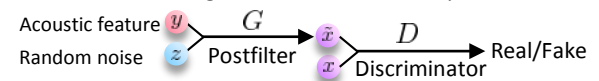
- Stochastic counterpart of Fujisaki model
- Parameter estimation using powerful statistical inference techniques

Phrase/accent command modeling with HMM

Allows intonation control

GAN-based postfiltering and conversion

- Train convertor G so that it can deceive discriminator D that tries to distinguish real data from samples from G



Allows generation of high-quality speech samples

Reference

- 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.
- 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