Vol.2019-MUS-123 No.10
Vol.2019-SLP-127 No.10
2019/6/23

音声合成のための信号処理・統計モデル

才野　慶二郎[1,a)]

概要：本チュートリアルでは，テキスト音声合成のタスクに用いられる主要な技術について概観する．テキスト音声合成タスクは典型的には (1) テキストから言語特徴量を得る，(2) 言語特徴量から音響特徴量を得る，(3) 音響特徴量から波形を得る，という 3 つの部分問題で構成される．中でも特に盛んに取り組まれているのが (2) と (3) である．(2) では，かつては音声素片を選択的に使用する単位選択型手法が主流であった一方で，近年は統計モデルを使用して尤もらしい音響特徴量を推論する統計的パラメトリック手法がよく用いられる．(3) ではボコーダと呼ばれる決定論的な信号処理手法が広く用いられるとともに，近年ではディープニューラルネットワークに直接波形を生成させる手法も用いられはじめている．また (1) と (2) を 1 つのモデルで一気に解くような End-to-end を指向した深層学習手法も近年提案され，注目を集めている．

An overview of technology for speech synthesis

[1] ヤマハ株式会社
10–1, Nakazawa–cho, Hamamatsu, Shizuoka 430–8650, Japan
[a)] keijiro.saino@music.yamaha.com