1-4-5
Blizzard Challenge 2018 のための NITechテキスト音声合成システム

◎沢田慶[1,2]，吉村建慶[1]，橋本佳[1]，
大浦圭一郎[1]，南角吉彦[1]，徳田恵一[1]

[1]名古屋工業大学 (NITech)
[2]マイクロソフトディベロップメント株式会社

日本音響学会2018年秋季研究発表会 2018年9月12日

はじめに

- **テキスト音声合成 (TTS) システム**
  - スマートフォン・スマートスピーカーの登場により利用拡大
  - 高音質・様々な発話スタイル・多言語の需要増大
  - 深層学習の導入により劇的な向上
    - *DNN, LSTM, WaveNet, Char2Wav, Deep Voice, Tacotron等*
- **TTSシステムの評価**
  - 学習コーパス・タスク・受聴試験が異なると直接比較が困難
  - Blizzard Challenge [Black et al. ’05]
    - *参加チームは共通の学習コーパスを用いてTTSシステムを構築*
    - *主催者による大規模主観評価実験によりTTSシステムを評価*
- **NITech TTSシステム**
  - 2005年から統計的音声合成システムを提出

Blizzard Challenge 2015−2018

- タスク
  - 児童書を児童に読み聞かせるのに最適なTTSシステムの構築
- データ
  - イギリス英語の女性話者1名により収録
    - *2015年 (パイロットタスク): 2時間*
    - *2016年: 5時間*
    - *2017, 2018年: 7時間*
  - 音声データとテキストに不一致が存在
    - *言い間違い，言い淀み，擬音語等*
  - 音声データには様々な発話スタイルが存在
    - *感情，キャラクター，歌声等*

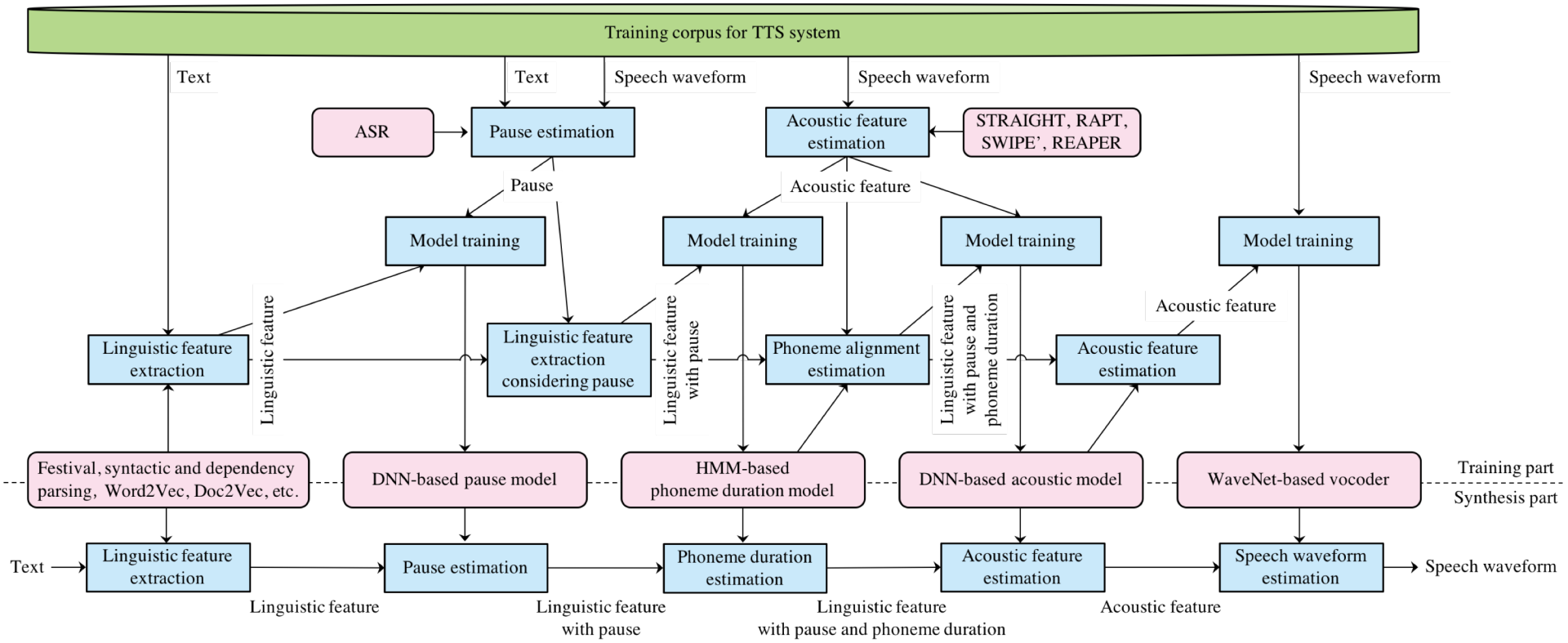
"I'm king of the jungle," roared Lion.
"I'm going to eat you all up."
"No!" cried the jungle animals.

Character1
Character2
Descriptive part

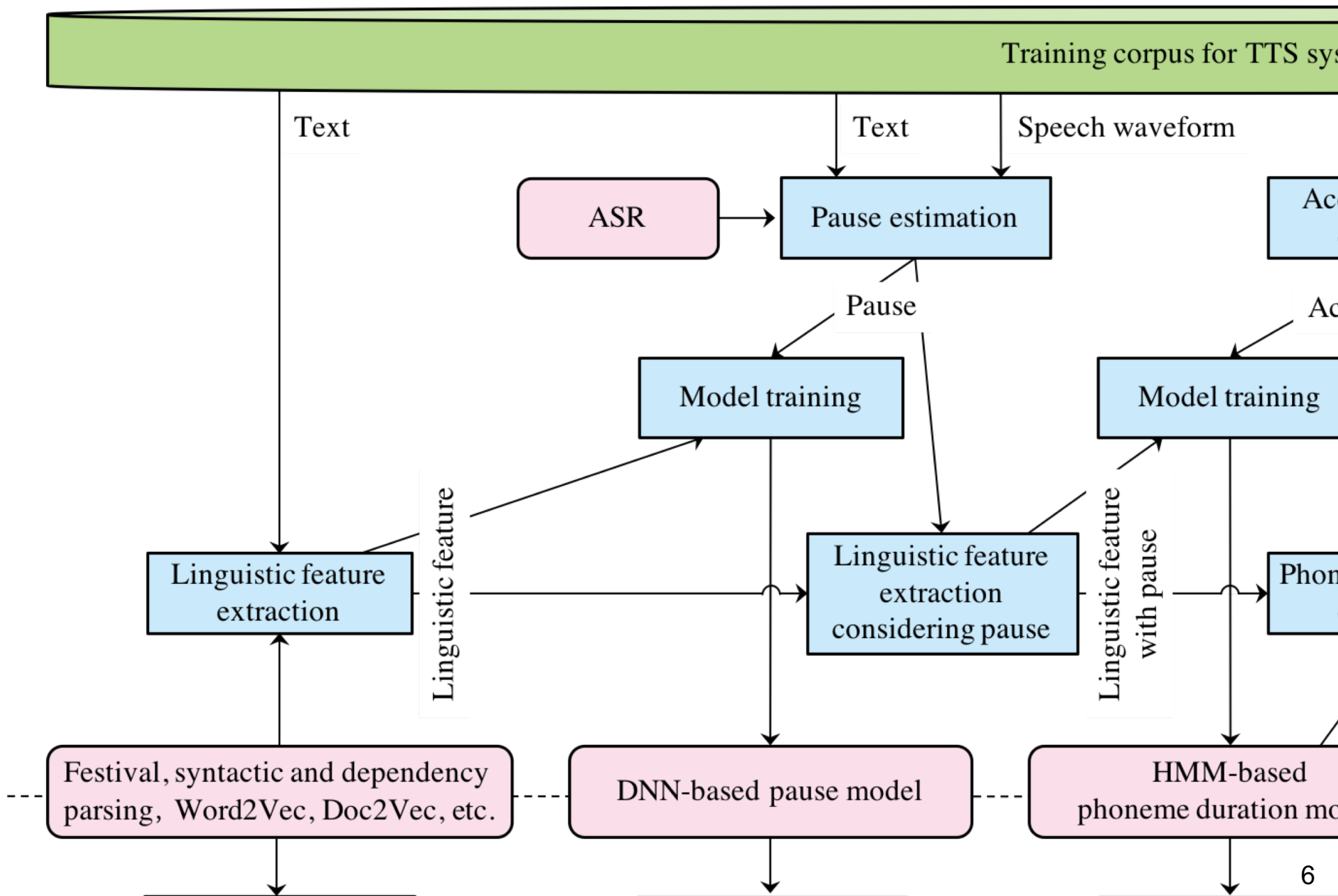
NITech 2015–2018 TTSシステム

- **NITech 2015 TTSシステム**
  - 音声認識を用いた学習データの取捨選択
  - 引用符に基づいた言語特徴量の導入
- **NITech 2016 TTSシステム**
  - 音声認識を用いた学習コーパスの自動構築
  - 構文解析に基づいた言語特徴量の導入
  - GV付きトラジェクトリ学習に基づくDNN音響モデルの導入
- **NITech 2017 TTSシステム**
  - 係り受け解析に基づいた言語特徴量の導入
  - 発話スタイルを予測し再現する言語特徴量の導入
  - GV付きトラジェクトリ学習に基づくMDN音響モデルの導入
- **NITech 2018 TTSシステム**
  - ポーズ挿入モデルの導入
  - WaveNetボコーダの導入

NITech 2018 TTS system



NITech 2018 TTS system

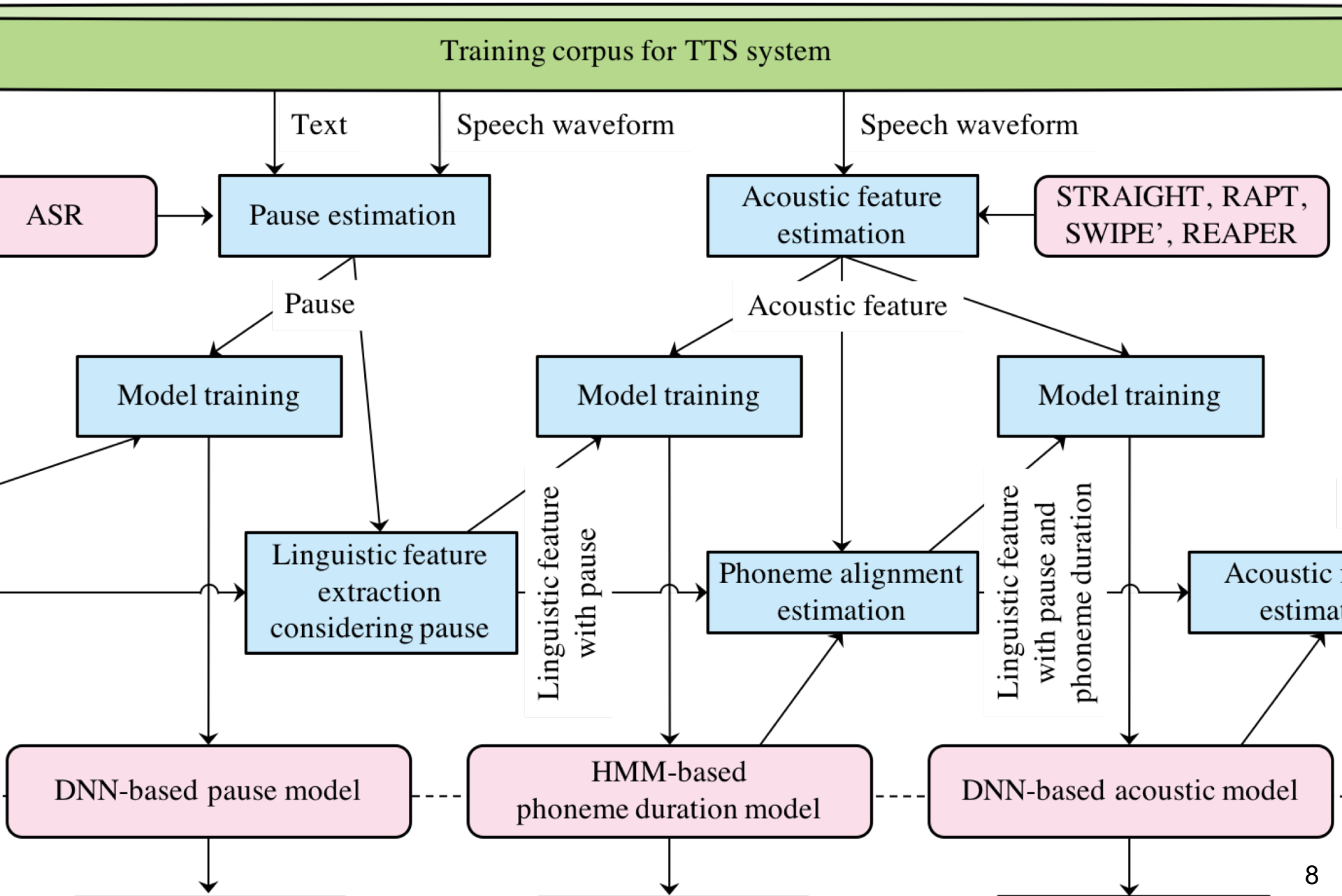


ポーズ挿入モデル

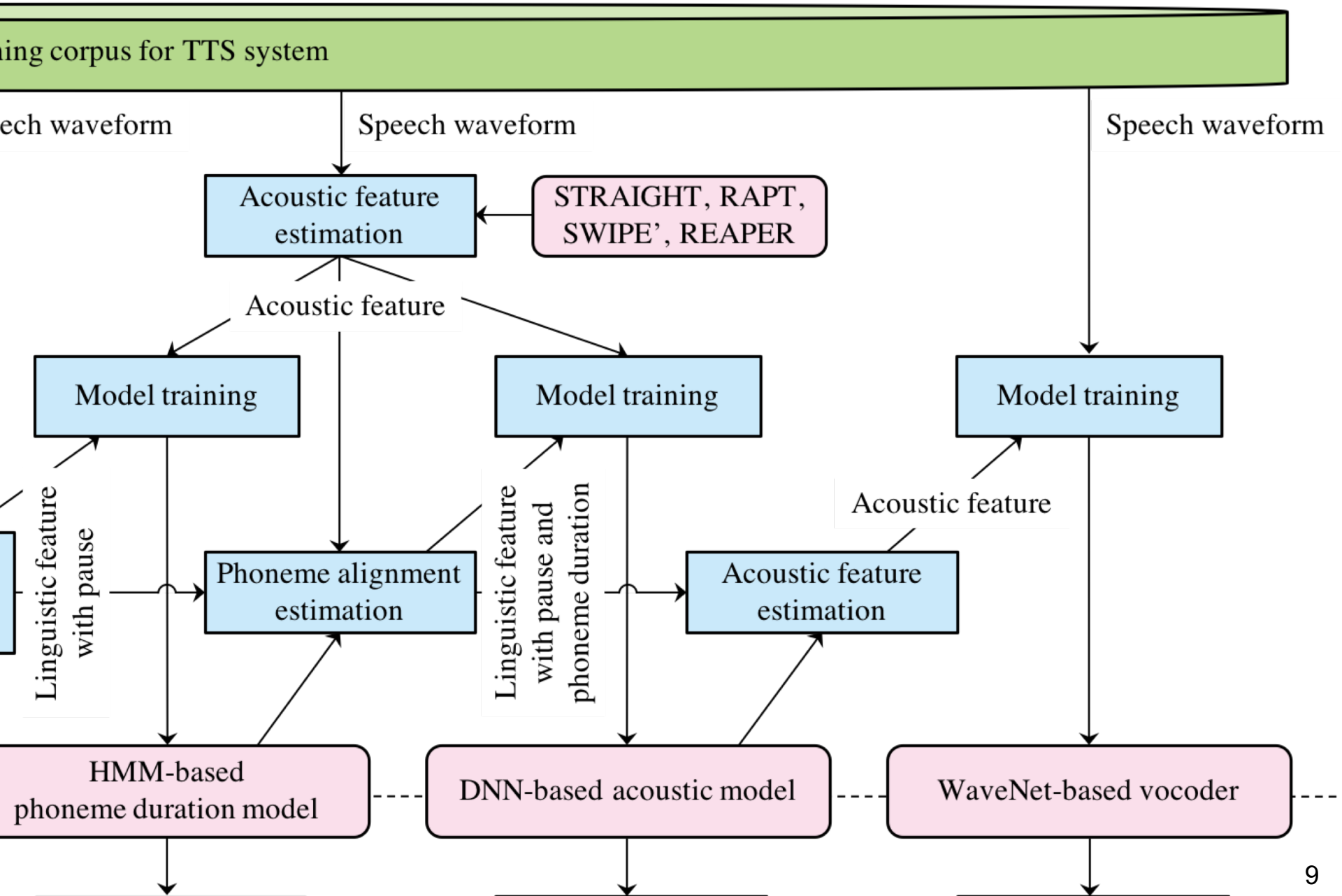
- **自然な合成音声には適切な位置にポーズが必要**
  - オーディオブックでは感情表現にポーズを利用
    ⇒ 学習コーパスのポーズ挿入スタイルを再現するモデル
- **学習コーパスのポーズ推定**
  - 全単語境界にショートポーズを含む音素アライメント推定
  - ショートポーズモデルにはHMMのスキップあり構造を利用
  - ショートポーズの継続長が閾値以上の場合にポーズ判定
- **ポーズ挿入モデル**
  - Bi-directional gated recurrent unit (GRU)
  - 入力：単語・文単位の言語特徴量
  - 出力：単語の後にポーズが挿入されるか否か (0 or 1)

学習コーパスのポーズ挿入スタイルを再現可能

NITech 2018 TTS system



NITech 2018 TTS system



WaveNet-based vocoder

- **フレーム単位のボコーダ**
  - 音響特徴量から容易に波形を生成
  - 波形生成時に合成音声の品質劣化
    ⇒ ニューラルボコーダの導入
- **WaveNetボコーダ** [van den Oord et al.; ’16], [Tamamori et al.; ’17]
  - 波形の直接モデル化・生成が可能
  - 分類問題として音声波形をモデル化
  - 白色の量子化ノイズが発生
    ⇒ ノイズシェーピング量子化を導入
- **メルケプストラムに基づくノイズシェーピング量子化**
  - 人間の聴覚特性を考慮した量子化ノイズ [Yoshimura et al.; ’18]
  - 音声符号化におけるポストフィルタリングを導入 [Tokuda et al.; ’94]

高品質な音声波形の生成を実現

実験条件 (1/3)

音声認識に基づく学習コーパス自動構築の条件

配布された学習コーパス	1258ページ
音響特徴量	12次元のMFCC + Δ + ΔΔ
音響モデル	3状態 left-to-right トライフォン GMM-HMM
言語モデル	トライグラムモデル
学習データの取捨選択閾値	90%
TTSのための学習コーパス	924ページ

ポーズ挿入モデルの条件

入力特徴量	251次元の言語特徴量
DNNのモデル構造	Bi-directional gated recurrent unit, 3層の中間層，128ユニット，ReLU
学習アルゴリズム	Adam，ドロップアウト率 20%

実験条件 (2/3)

継続長モデルの条件

サンプリング周波数	32 kHz
音響特徴量	63次元STRAIGHTメルケプ，対数基本周波数， 32次元のメルケプ非周期成分 + Δ + ΔΔ
クラスタリングの質問数	925質問
HMMのモデル構造	5状態 left-to-right MSD-HSMM

音響モデルの条件

サンプリング周波数	32 kHz
音響特徴量	63次元STRAIGHTメルケプ，対数基本周波数， 有声無声情報，32次元のメルケプ非周期成分
言語特徴量	1685次元
DNNのモデル構造	Single-mixture density network, 3層の中間層，8000ユニット，シグモイド
学習アルゴリズム	SGD，ドロップアウト率 60%, GV付きトラジェクトリ学習

実験条件 (3/3)

WaveNetボコーダの条件

サンプリング周波数	32 kHz
量子化	8ビットμ-law
ノイズシェーピングパラメータ	γ=0.1, β=0.1
WaveNetのモデル構造	Dilation 1, 2, 4, ..., 512 を3段, dilation, residual, skipは256チャンネル
WaveNetの補助特徴量	音響モデルから生成した98次元の音響特徴量
学習アルゴリズム	Adam

デモ

画像は Usborne Publishing より引用

文単位の実験結果

Naturalness

MOS	ID
4.8	A
4.0	K
3.7	J
3.5	I
3.0	L, M
2.9	B
2.8	D
	⋮

Similarity

MOS	ID
4.5	A
3.9	K
3.6	J
3.5	I
3.4	L
3.2	B
3.0	M
	⋮

Intelligibility

WER	ID
11	I
14	E, O
15	D, G
16	K
17	N
18	J
20	F
	⋮

A: 自然発話
B, C, D, E: ベンチマークTTSシステム
I: NITech TTSシステム
Red line: NITech TTSシステムと有意な差があるTTSシステム

ページ単位の実験結果

Overall impression

MOS	ID
48	A
38	K
34	J, I
29	B
28	L
	⋮

Pleasantness

MOS	ID
48	A
37	K
33	J, I
28	L, B
26	M
	⋮

Speech pause

MOS	ID
48	A
36	K, J
32	I
31	D, G
30	E
	⋮

Stress

MOS	ID
48	A
36	K
35	J
33	I
30	D, G
	⋮

Intonation

MOS	ID
48	A
37	K
35	J
33	I
28	D
	⋮

Emotion

MOS	ID
48	A
38	K
35	J, I
31	B
30	M
	⋮

Listening effort

MOS	ID
49	A
37	K
34	J
33	I
28	D
	⋮

NITech 2017 と 2018 TTSシステムの比較

- **WaveNetボコーダの導入**
  - 自然性と話者類似性が向上
  - 発音が一部曖昧になる合成音声
    - *複数のコーディックやノイズがWaveNetの学習を困難に*
  - 発話スタイルの再現性が低下
    - *様々な発話スタイルを再現するWaveNet学習にはデータ不足*

2017	2018

むすび

- **Blizzard Challenge 2018のためのNITech TTSシステム**
  - ポーズ挿入モデルの導入
  - WaveNetボコーダの導入
  - 大規模受聴試験の結果
    - *自然性・話者類似性・明瞭性で高いスコアを達成*
  - NITech 2017 と 2018 TTSシステムの比較
    - *自然性・話者類似性が向上*
    - *WaveNetボコーダの精度が不十分*

- **今後の課題**
  - ニューラルボコーダで様々な発話スタイルを合成
  - End-to-endアプローチの導入