2-Q-21 Blizzard Challenge 2017のためのNITechテキスト音声合成システム

◎沢田慶, 橋本佳, 大浦圭一郎, 南角吉彦, 徳田恵一 (名工大; NITech)

日本音響学会2017年秋季研究発表会
2017年9月26日

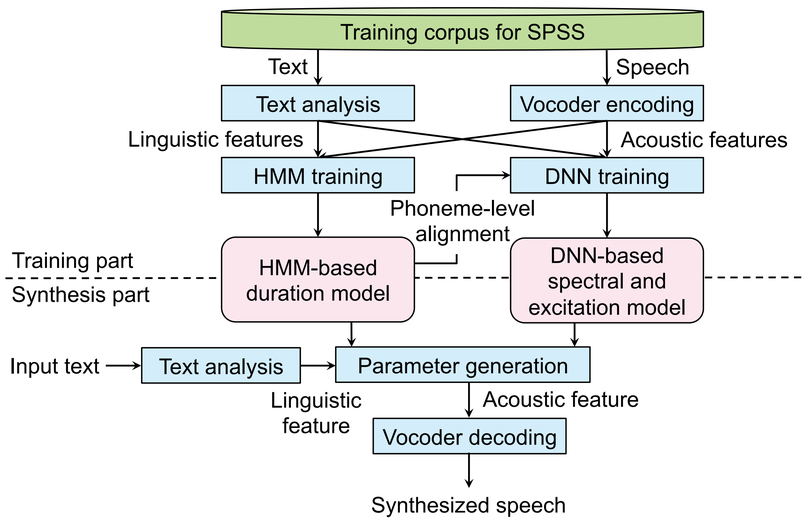
Blizzard Challenge 2017

- **Blizzard Challenge [Black et al. '05]**
  - 同一の学習コーパスを用いてTTSシステムを構築
  - 大規模主観評価実験によるTTSシステムの評価
- **Blizzard Challenge 2017**
  - データ
    - *英語の児童書のオーディオブック56冊 (7時間) が学習コーパスとして配布*
    - *イギリス英語の女性話者1名*
    - *音声データとテキストには不一致が存在*
    - *音声データには様々な発話スタイル・感情・キャラクターが存在*

"I'm king of the jungle," roared Lion.	キャラクター1
"I'm going to eat you all up."	キャラクター2
"No!" cried the jungle animals.	地の文

  - 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*

NITechテキスト音声合成システム



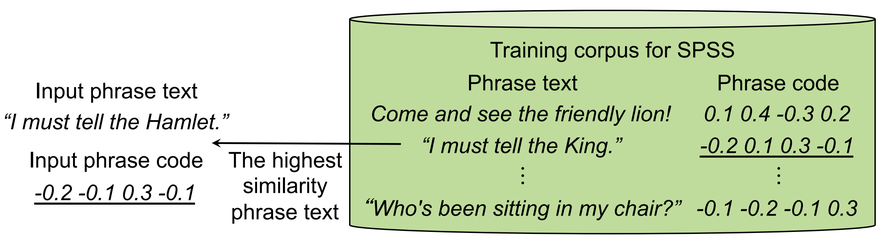
① 音声認識を用いた学習コーパスの自動構築
② オーディオブックのための言語特徴量の設計
③ 系列内変動 (GV) を考慮したトラジェクトリ学習を用いた密度ネットワーク (DN)

① 音声認識を用いた学習コーパスの自動構築

- **音声認識性能の改善**
  - ブック適応言語モデルと話者適応音響モデル
    ⇒ オーディオブックのデータを用いて音声認識性能改善
- **学習データの選択**
  - ブックテキストと認識テキストの単語一致率
    - *単語一致率が高 ⇒ 正解テキストと類似する可能性大*
  - 単語一致率に閾値を設けたデータ選択
    ⇒ 適切な閾値の利用により学習コーパスの品質向上
- **テキストの改善**
  - ブックテキストと認識テキストを用いた単語ネットワークを構築
  - 単語ネットワークに制限した音声認識
    ⇒ 正解テキストと類似するテキストの推定

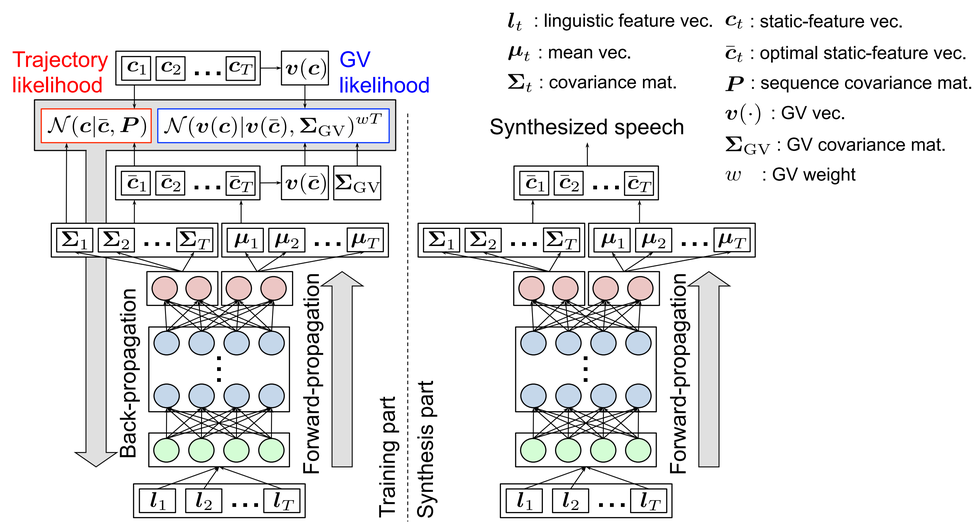
② オーディオブックのための言語特徴の設計

- **言語特徴量**
  - テキストから得られる発音を表す特徴量
  - 学習コーパスに適した設計が重要
- **新たに導入した言語特徴量**
  - ページ単位の情報
    - *文単位以上の情報を利用*
  - 構文解析と係り受け解析の情報
    - *文の構造の情報を利用*
  - 文の種類
    - *文の種類による発話スタイルの区別*
  - 引用符の情報
    - *地の文と会話文による発話スタイルの区別*
  - 単語コード
    - *各単語の区別*
  - フレーズコード
    - *各フレーズの区別と発話スタイルの再現*
- **フレーズ単位の言語特徴量**
  - doc2vecを言語特徴量として直接利用
    ⇒ テキストと音声のスタイル不一致による平均化
- **フレーズコード**
  - 学習部
    - *全てのフレーズに一意のコードを割り当て*
    - *固定次元の標準正規分布に基づく乱数ベクトル*
      ⇒ *平均化を回避し各フレーズのスタイルをモデル化*
  - 合成部
    - *学習・テストフレーズをdoc2vecによりベクトル化*
    - *学習・テストフレーズの類似度をdoc2vecで計算*
    - *最も類似度が高い学習フレーズのコードを利用*



③ GVを考慮したトラジェクトリ学習を用いたDN

- **密度ネットワーク (DN)**
  - DNNの出力が単一正規分布のパラメータ
- **系列内変動 (GV) を考慮したトラジェクトリ学習**
  - トラジェクトリ学習 ⇒ 系列単位の最適化
  - GVを考慮 ⇒ 過度な平滑化を抑制した最適化



大規模主観評価実験

学習コーパス自動構築の条件

児童書	56冊, 1258ページ
標本化周波数	16 kHz
フレーム	長さ: 25 ms, シフト: 10 ms
音響特徴量	12次元 MFCC + Δ + ΔΔ
音響モデル	3-state left-to-right GMM-HMM
言語モデル	tri-gram
単語一致率閾値	単語一致率: 90%
学習コーパス	921ページ

音声サンプル

TTSシステム構築の条件

標本化周波数	44.1 kHz
フレーム	シフト: 5 ms
HMMの構造	5-state left-to-right MSD-HSMM
音響特徴量 (HMM)	49次元STRAIGHTメルケプ, 24次元メルケプ非周期成分, 対数基本周波数, 上記のΔとΔΔ
質問の数	925の質問
MDNの構造	3層の隠層, 8000ユニット, 活性化関数: シグモイド, ドロップアウト率: 60%, GV重み: 0.001
音響特徴量 (MDN)	69次元STRAIGHTメルケプ, 34次元メルケプ非周期成分, 補間対数基本周波数, 有声無声情報
言語特徴量	925次元の言語特徴量, 10次元の継続長特徴, 150次元の単語コード, 600次元のフレーズコード

実験結果

- システム
  - *A: 自然音声*
  - *B: Unitベンチマーク*
  - *C: HMMベンチマーク*
  - *D: DNNベンチマーク*
  - *E-Q: 参加者のシステム*
  - ***L: NITechシステム***
- 評価基準
  - *OI: overall impression*
  - *PL: pleasantness*
  - *SP: speech pauses*
  - *ST: stress*
  - *IN: intonation*
  - *EM: emotion*
  - *LE: listening effort*
  - *NAT: naturalness*
  - *SIM: speaker similarity*
  - *WER: word error rate*

システム	ページ単位 60段階MOS							文単位 5段階MOS		無意味文
	OI	PL	SP	ST	IN	EM	LE	NAT	SIM	WER
A (natural speech)	47	46	47	47	47	47	47	4.7	4.5	-
I (hybrid)	38	38	36	36	37	38	36	4.0	4.0	34
G (neural vocoder)	32	31	34	33	33	34	32	3.6	3.1	35
L (SPSS)	**31**	**30**	**31**	**31**	**31**	**33**	**31**	**3.6**	**3.0**	**30**
E (hybrid)	31	32	27	27	28	32	28	3.5	3.9	38
P (hybrid)	31	31	31	32	32	34	30	3.4	3.6	38
B (unit selection)	27	27	25	26	27	30	25	3.3	3.7	42
M (SPSS)	26	25	24	27	26	26	25	3.2	3.1	33
K (hybrid)	26	27	26	25	26	28	25	3.1	2.9	42
Q (unit selection)	25	26	23	23	24	28	23	3.1	2.8	44
D (SPSS)	25	23	30	28	26	24	26	2.7	2.5	30
H (clustergene)	24	24	25	25	23	21	24	2.7	2.1	39
J (SPSS)	22	21	24	23	22	20	22	2.7	2.6	35
F (SPSS)	21	22	24	24	24	25	21	2.6	2.4	46
C (SPSS)	16	15	21	19	18	18	17	1.7	1.9	38
O (SPSS)	11	11	22	19	16	16	11	1.5	1.5	60
N (SPSS)	8	8	16	13	12	11	7	1.1	1.2	85