過去の音声資料を用いた効率的な音声合成システムの構築に関する研究

名古屋工業大学・助教　酒向慎司

- 自分の声とは誰のものか？
- 声の本人らしさとは？

本発表の概要

- 隠れマルコフモデルによる音声合成手法
  - 基本的な原理
  - 音声合成モデル構築方法
- 故人の声合成の取組み
  - 概要と関連事例
  - 音声合成モデル作成
- まとめと課題

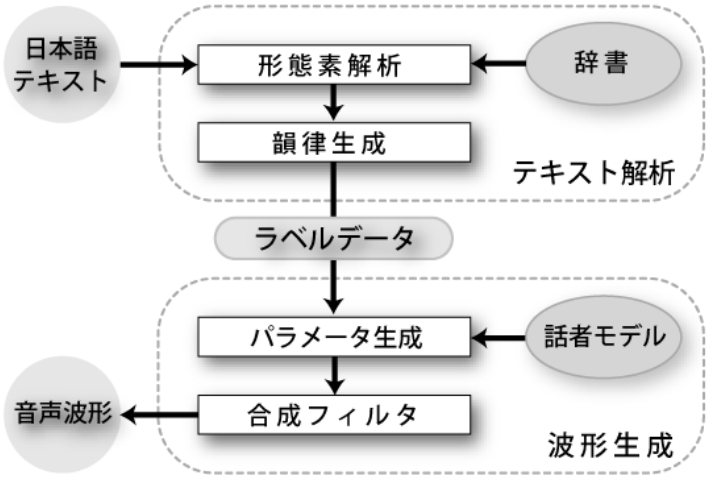
HMM音声合成技術

- テキストから音声を合成する技術
- 隠れマルコフモデル(hidden Markov model)による音声のモデル化
  - 音声波形の接続では無く音声の生成機構をモデル化
- 音声データから本人らしさを備えた音声合成
  - 一定量の音声データ＋言語情報が必要
  - 学習データからモデルの自動構築が可能

  - 声の品質はデータベースに依存

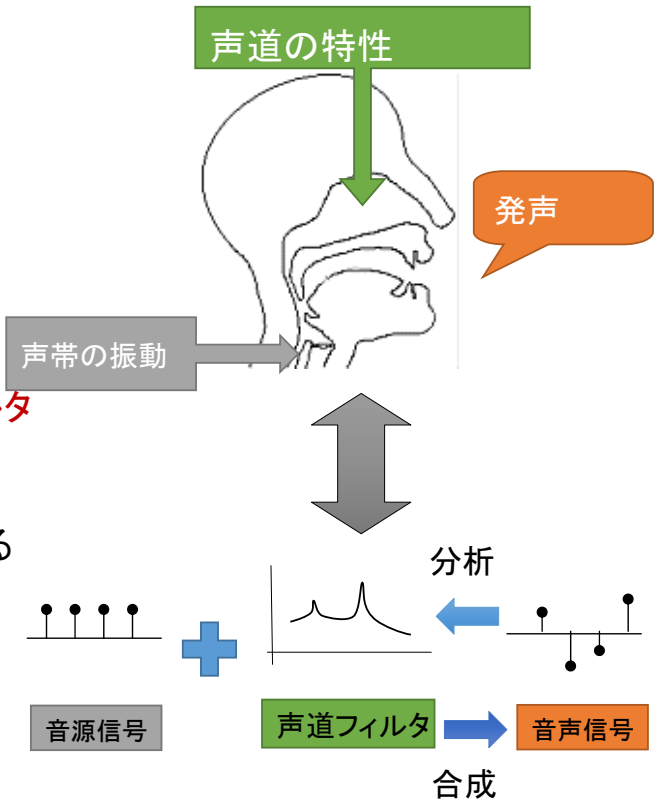
音声合成処理のフロー

1. 言語解析（形態素解析）
2. 音韻交代・チャンキング
3. 波形生成　←HMM音声合成



ソース・フィルタモデル

- 音声生成の線形モデル
  - 音源信号：声帯の振動（ピッチ）
  - 声道フィルタ：声道の形状
- 音声分析合成モデル
  - 分析・合成フィルタ
  - LPC, PARCORなどの分析合成系
  - メルケプストラム分析・MLSAフィルタ
- 音声の特徴抽出の基礎
  - フィルタ係数：声道の特徴を有する
  - 音声認識に活用
  - フィルタ係数と音源信号から音声を復元することができる



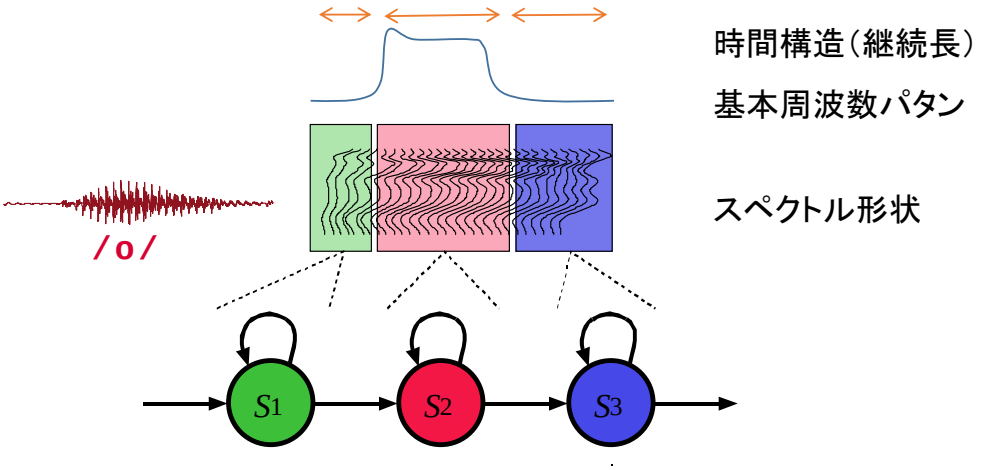
HMMの学習

- HMM音声合成では何を獲得(学習)するか？

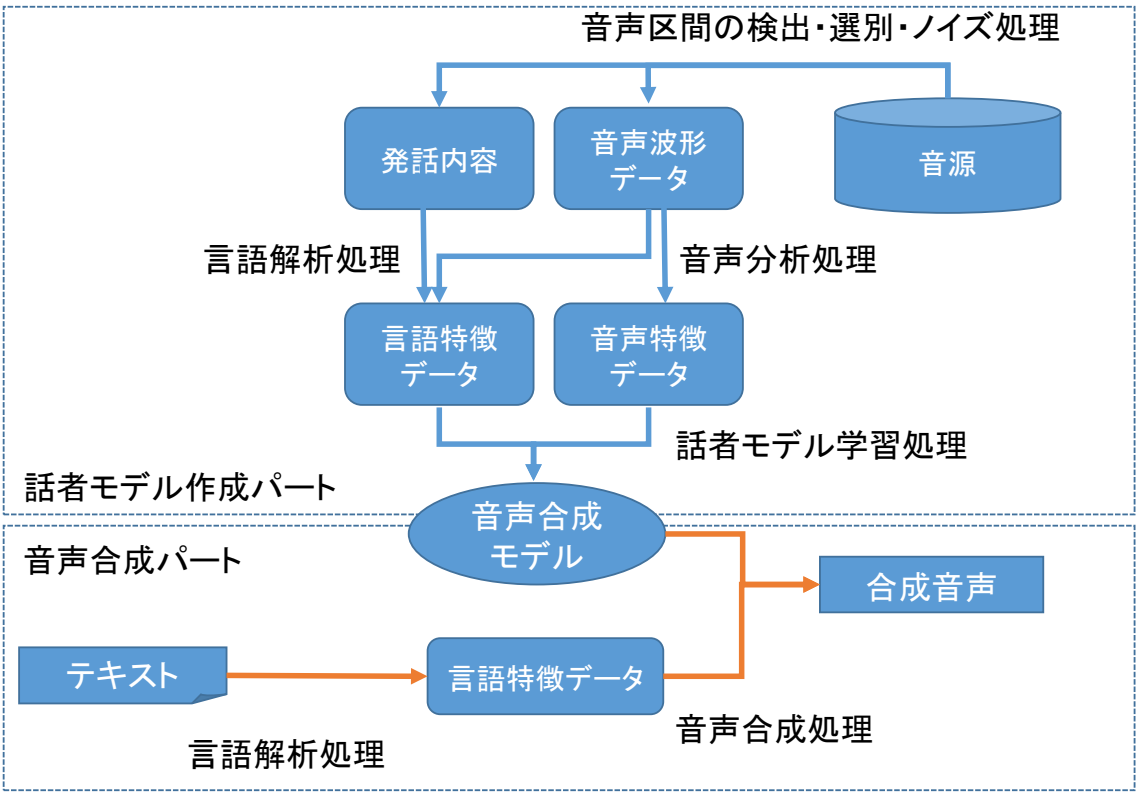
⇒特定の音素波形を作る(再現する)ための材料

- 音のスペクトルパタンの形状
- ピッチの形状
- それらの時間的な変化

分析合成系のパラメータを決定することで音声を合成可能

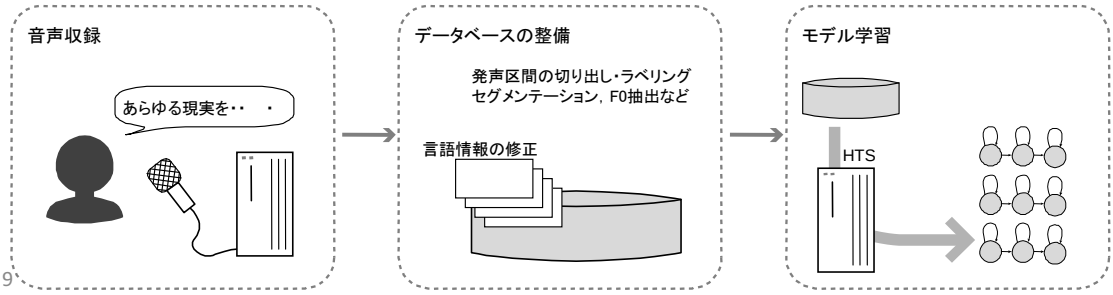


処理全体の流れ



標準的なモデル学習の手順

- 音声収録
  - 音韻バランス文章など（例：ATR音韻バランス文503文章）
- データベースの整備
  - 音素セグメンテーション
  - 基本周波数(F_0) 抽出
  - 言語・韻律ラベリング
- モデル学習
  - モデル学習ツールキット(HTS)を利用



学習サンプル数による違い

- 文章数が増えるほど
  - 雑音が低減しクリアな音質に
  - 韻律がより正しく再現される

25	50	100	250	503	原音声

故人の声合成の取組みの概要

- 目的：故人の父親の声を再現して、娘の結婚式にお祝いのメッセージを音声合成する
- 素材（音源データ）：約13時間のホームビデオから本人の音声を利用
  - 本人音声の自動抽出・手動処理
  - 本人音声モデルの学習
- 完成したメッセージを披露宴にて披露

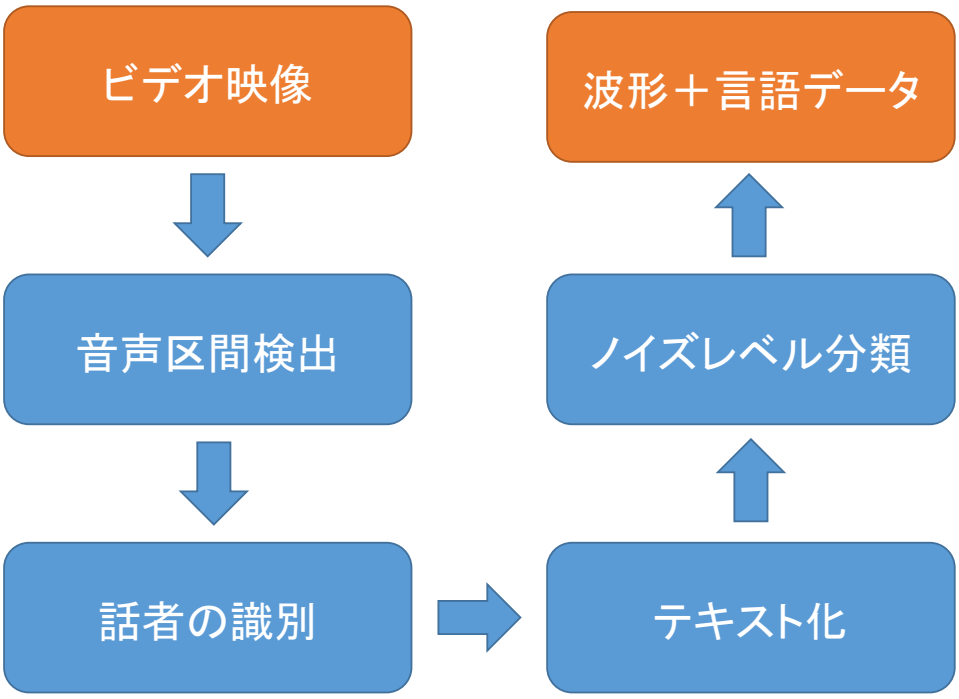
類似した事例（国内）

- VOCALOID「ウエキロイド」 (2010年ごろ)
  - YAMAHA株式会社の研究チームが植木 等さんの音源データをもとに試験的に作成
- 美空ひばりさん (2005年ごろ)
  - 日本音響研究所が作成し、長男の披露宴で祝辞を再生
- VoiceBank (2014年～）
  - 国立情報学研究所による取組み
  - 声帯摘出者やALS患者などに本人の声を合成
- マイボイス（2013年～）
  - 慶応義塾大学による取組み
  - ALS患者が本人の声を合成
  - 意思伝達装置と組合せた活用事例多数

音声素材について

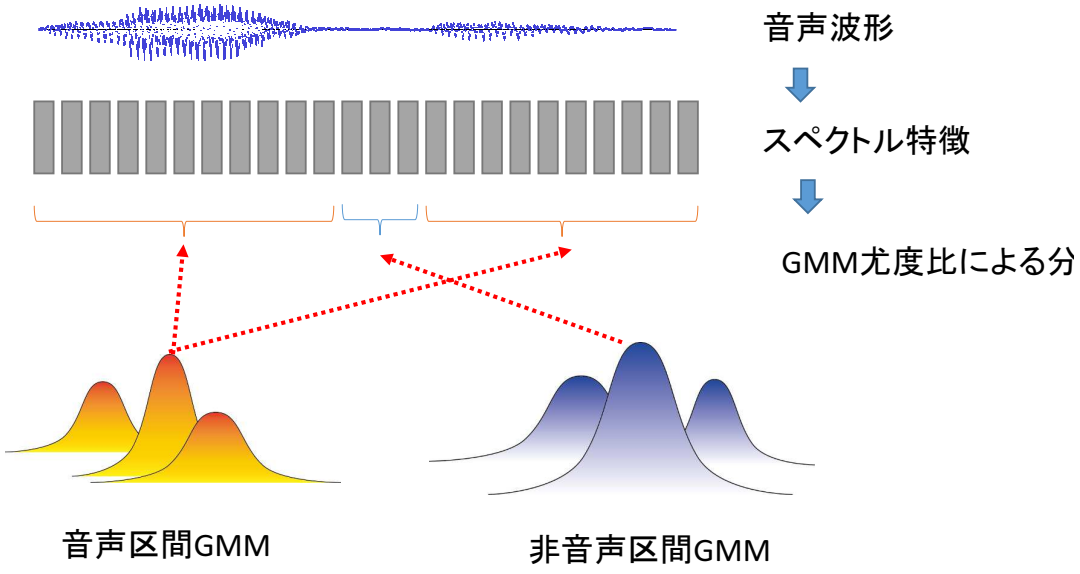
- 合計約13時間23分の音声データ
  - 家庭内・旅先などのホームビデオ
  - VHS×4、Hi8×2、MiniDV×5
  - 長期間に渡り本人の声質も変化
  - 録音環境も様々
- 約20分の故人の声の学習用データを抽出
  - 比較的録音状況の良いminiDVから使用

処理手順の概要



自動発話区間検出(VAD)

- 混合ガウスモデル(GMM)による音声・非音声区間の自動検出



話者の自動分類とテキスト化

- 一般的な話者識別手法
  - GMMによる音声特徴のモデル化
  - i-vector（因子分析により話者固有の特徴を表現）
    - 映像データ中の音声はi-vectorによる手法に不向き
  - 録音媒体ごとに少量データの本人音声でGMMを学習し、自動分類
- テキスト化
  - 音声認識エンジンによるテキスト化
  - 短い会話文のため自動化は困難

抽出された声データの例

01-001:ここはどこですか？よっちゃん。
01-002:はい。
01-003:そう、今とってるよ。
01-004:いいよー。
01-005:木の上へ置いてきな。
01-006:ん？。ほぉー。
01-007:焼くほどでもなんでも関係ないか。ね。
01-008:これは、ホタテガイ。
01-009:うし！。じゃあ、これは、サンゴにみえる。これ。

音響モデルの学習

- 区間検出された音声波形:約8000秒≒2時間
  - 検出誤り
  - ノイズの混入 → 合成品質の低下
  - 録音環境の差異
  - テキスト化の誤り
- テキスト化はほぼ手動修正
- ノイズレベルの差により分類し合成品質の良くなるクラスを調整→20分

名古屋TV「UP!」にて特集

まとめ

- 音声データ収集には自動処理には限界があり手動に頼る部分も大きかった
- 品質はそれほど高いものでは無かった
  - 本人らしさを感じるのは声質よりも話し方によるところも大きい
  - 話し方の定量的な表現は難しく
- 他の依頼を受けて実施したが同じレベルの合成音は達成できていない

謝辞

本研究の実施にあたり下記の方々のご協力・ご支援を頂きました。御礼申し上げます。

- データ提供者（ご遺族の方々）
- 竹田設計工業株式会社
- 財団法人人工知能研究振興財団