NINJAL
National Institute for Japanese Language and Linguistics

音声言語資源構築の諸問題

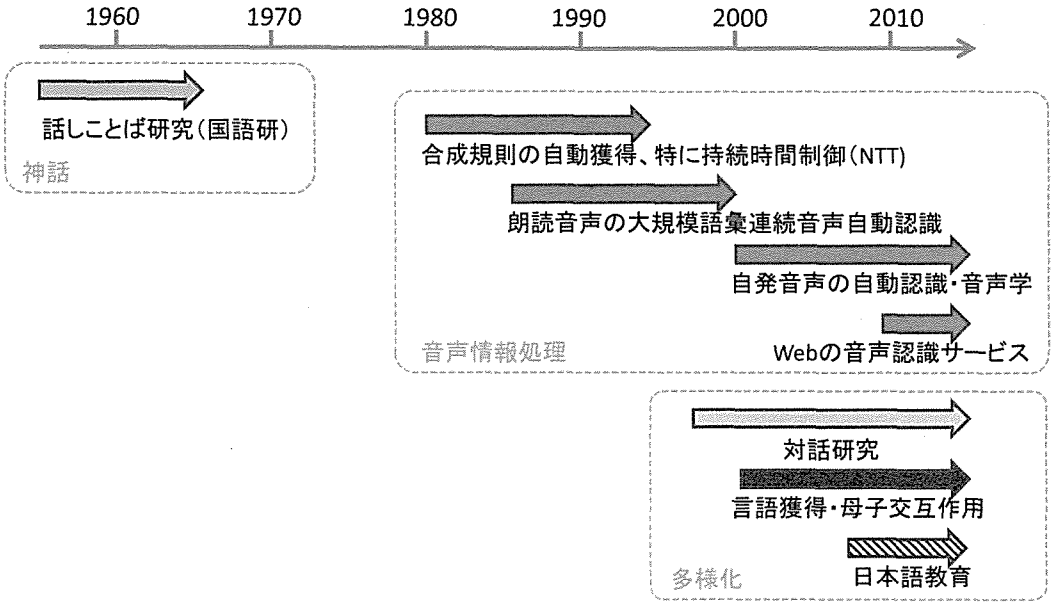
前川 喜久雄
国立国語研究所 言語資源研究系

大学共同利用機関法人 人間文化研究機構
国立国語研究所
National Institute for Japanese Language and Linguistics

内 容

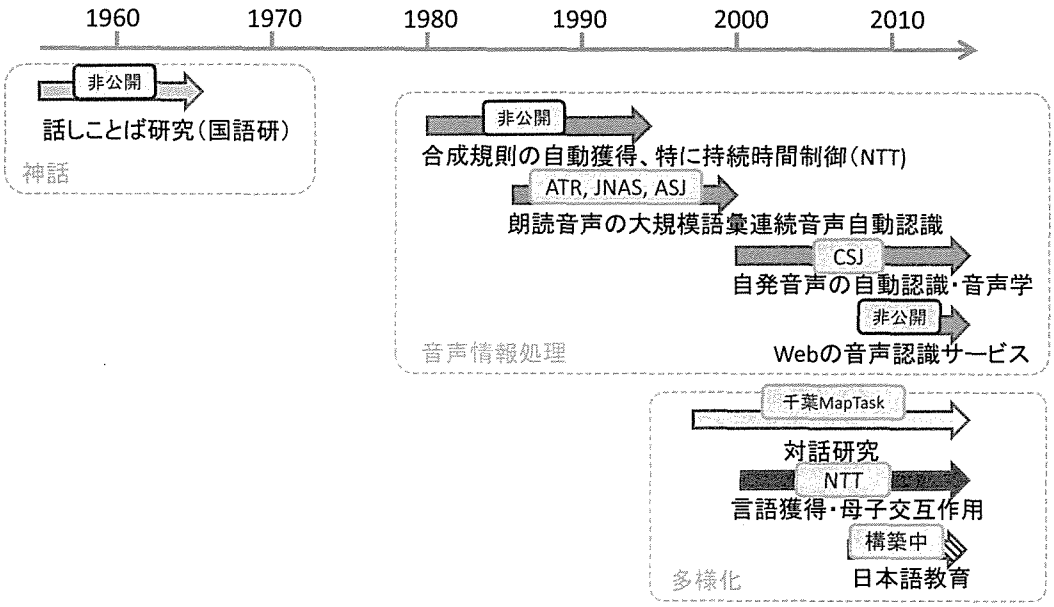
- 音声言語資源開発の流れ
  - 時代区分
  - 代表的コーパス
  - 規模
  - アノテーション
  - 多様化の時代の特徴
- 現時点の技術的な課題
  - レジスターの拡大
  - パラ言語情報
  - マルチメディア化
  - 自動化
  - 著作権処理
  - 共有化

時代区分



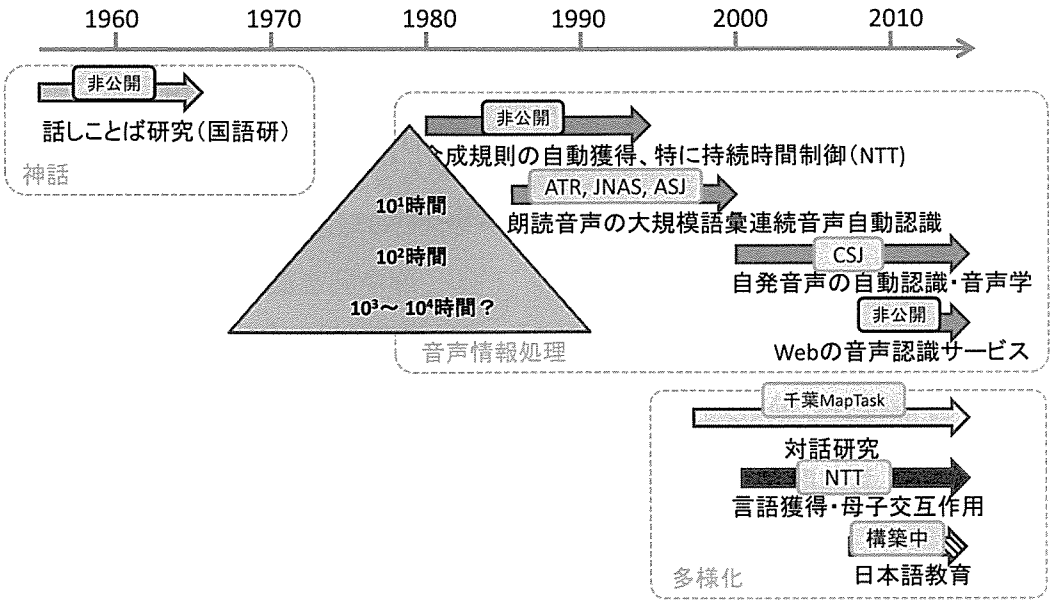
NINJAL National Institute for Japanese Language and Linguistics

代表的コーパス

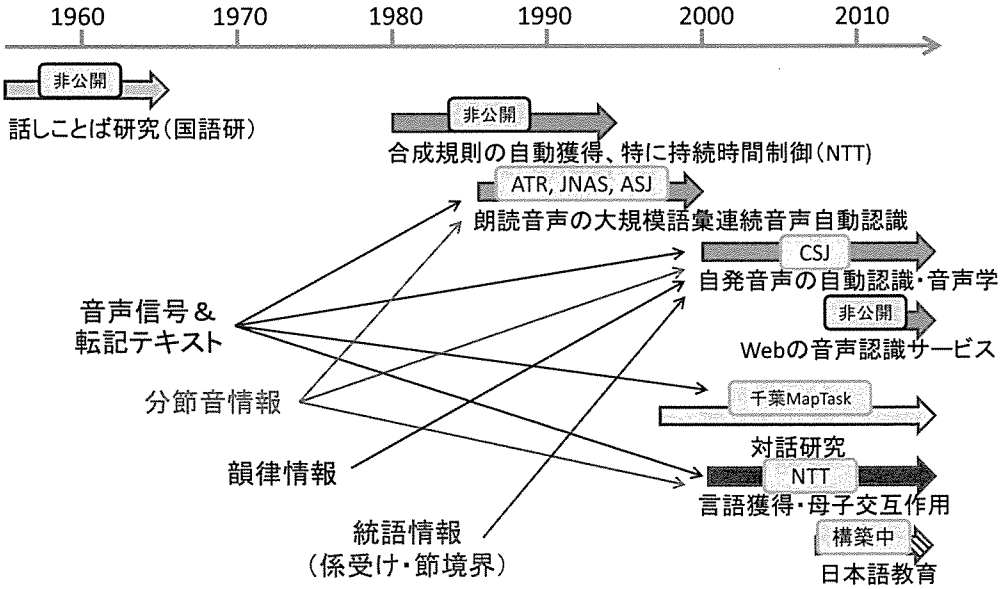


NINJAL National Institute for Japanese Language and Linguistics

規模の拡大



公開コーパスのアノテーション



多様化の時代の特徴

■ 音声情報処理主導の時代
- 研究の方法論が明確(統計的モデリング)
- 単純(朗読音声)から多様(自発音声)へ
- 規模の拡大で多様性に対処
- コミュニティでの共同開発が有効に機能

■ 多様化の時代
- 方法論はバラバラ。領域によっては模索中
  「革袋は新しいが、新しい酒がない」状態
- しかしデータを重視する立場だけは共通
  いわゆる「理論的」研究への失望
- 共同開発が難しい時代?

内 容

■ 音声言語資源開発の流れ
- 時代区分
- 代表的コーパス
- 規模
- アノテーション
- 多様化の時代の特徴

■ 現時点の技術的な課題
- レジスターの拡大
- パラ言語情報
- マルチメディア化
- 自動化
- 著作権処理
- 共有化

レジスターの拡大

Register(使用域):誰がどのような場で誰に対してどのような目的で話すか?

- 朗読はひとつのレジスター
- CSJは複数の独話レジスターを収録(学会講演と模擬講演)
- 対話は課題志向中心
- 欠けているのは、
  自発的な対話(多人数対話を含む)
  職業的発話(教師, 政治家, 司会者, 車掌, 看護師, 母親 etc.)
  演劇 etc.

問題: いったいどれだけのレジスターがあるのか? どれが重要か。
調査が必要。

パラ言語情報

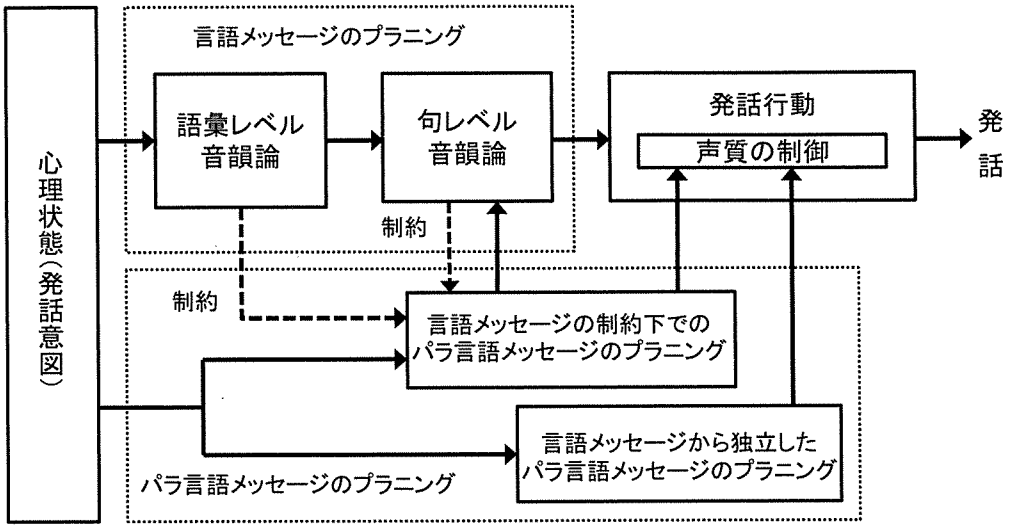
パラ言語情報(広義):意図・態度・意志的に表出された感情

- 音声コミュニケーションの必須要素(文字からは脱落する情報)
- 音声のプラニング段階で参照される情報(音韻論にも必要)

問題: 体系化がなされていない
関連する音声(音響)特徴の研究が不十分
演劇音声の分析などが有効と思われるがデータが存在しない
意図的な表出を広範囲に収集する試みがある(「千の声」)

パラ言語を含む音声生成のモデル

森・前川・粕谷『音声は何を伝えているか』(近刊)第3章より



パラ言語による声質制御の例

フィラー(filled pause)

- 言語的な意味は希薄(「不要語」)
- 談話管理上の機能は認められる

といわれてきているが、以下の3種の「エー」は同じか？

- ➢ 普通の「エー」
- ➢ Creaky な「エー」
- ➢ Breathyな「エー」

特に普通とcreakyの差は何らかの意図の差を反映していないか？

マルチメディア化

音声コミュニケーションは音声だけで行われるわけではない。
目は口ほどにものを言っている。

- 身振り
- 視線
- 表情
- 話し手と聞き手の位置関係

etc.

問題：　アノテーション体系が整備されていない
　　何が関与的特徴であるかについての研究の蓄積が不十分

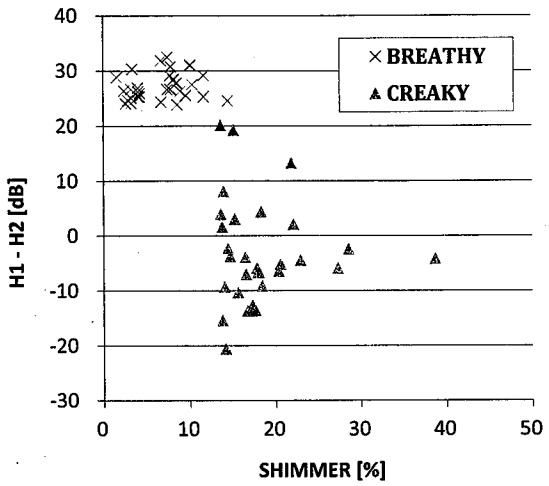
自動化

大規模なコーパス開発では、音声・映像データもアノテーションもできるだけ自動化したい。コストの問題あるが、客観的にみると人間よりもコンピュータの方が仕事の質が高いことも少なくない。自動化の現状は、

- 著作権問題を無視すればデータ収集の自動化は可能
- 形態論情報と統語情報はかなりの程度まで自動化可能
- 転記テキストの作成と分節音情報のラベリングは、単純な音声ならばある程度可能
- 韻律情報アノテーションの自動化は困難

※自動化の技術開発は非常に重要。できるものから自動化していく

自動化の一例：フィラーの発声様式



音声信号のshimmerとスペクトルの第1倍音と第2倍音の振幅差の計算で、息漏れ発声と軋み発声はかなりの程度まで分離できる（CSJコアのデータ、各軸とも上位30サンプル）。赤い△は Creaky-Breathy phonation

著作権処理

- 書き言葉（『現代日本語書き言葉均衡コーパス』）にくらべれば、『日本語話し言葉コーパス』の著作権処理は簡単だった。
  - 学会講演話者には事前に連絡して許諾を得た
  - 模擬講演話者は人材派遣会社を通して許諾を得た
  - 公開時には音声を含めて個人情報を無効化した

- 著作権処理に失敗して収録・公開を断念したのは、
  - NHKの報道番組「明日を読む」
  - 放送大学（ラジオ）の講義

※放送番組は権利のかたまりだが、放送局（少なくともNHK)は権利関係を管理していない。著作物の二次利用に関しては経営判断もある。

著作権処理

- 現状の大きな問題点は、苦労せずに処理できるはずなのに、処理を行っていないデータがあること(次項参照)。
- 著作権に関する国民の理解は十分なものではない(人格権と財産権の違いを知る人はまずいない)。著作権を利用する立場にある研究者は、誤解をなくすための努力がもとめられる。
- 個人相手でも、法人相手でも、研究の社会的価値をはっきり説明できなければ、許諾はもらえない。
- 著作権法の改正も必要(fair use条項の実現)

共有化

- 共有化の哲学: データをすべて自力で作ることはできない。他人のデータを利用させてもらい、自分のデータは公開しよう。
- 税金による研究に伴う義務: 広義の税金(科研費、各種法人の運営費交付金等)によって作った大規模データは、研究終了後は公開するのが原則。そのために必要な最低限の著作権処理はデータ作成者の義務。
- 研究の信憑性を高めるためにも公開が有効。
- 共有化のノウハウ: 文書化ノウハウの共有が重要。

共有化のための組織

■ やはり何らかの組織が必要。私見としては、日本の場合、今はまだ国立機関の関与が望ましい。なぜなら、

- 日本語の市場規模が小さすぎてLDCの真似はできない
- 著作権処理などで信頼を得やすい
- ひもつきになりにくい
- つぶれにくい（最近はかなり怪しいが）

■ 反対に国立組織の問題は、

- 行政の縄張り意識（文科省 vs.総務省、省庁内部の競争）
- 金銭のあつかいが極端に不便
- 政治（家）の介入

まとめとしての提言

- 多様化の時代には、ほっておけば交流がなくなる

  ⇒ 他領域の研究者が何をしようとしているかを知ることが困難

  ⇒ データ作成の技術とノウハウを共有することが困難

⇩

- 研究者の交流の場が重要
- 言語資源に特化した研究会を作れないか
- 言語資源の開発を業績として認知させるための努力も必要