

# 授業アーカイブの翻訳字幕 自動作成システムの試作

須藤 克仁 林 輝昭 西村 優汰\* 中村 哲

奈良先端科学技術大学院大学

先端科学技術研究科 先端科学技術専攻 情報科学領域

\*在学中の研究

Augmented Human Communication Laboratory

NAIST®

本研究の一部はJSPS科研費JP17H06101「次世代音声翻訳の研究」の助成を受けたものです。

# 概要

- 授業映像に字幕を付けます

The image shows a video player interface for a lecture. The main content is a slide titled "音声の知覚" (Auditory Perception) with a bulleted list: "明瞭度と了解度", "知覚単位と文脈", and "カテゴリー知覚". A yellow callout box with a dashed border contains the text: "そういうこう。パターンがあってでそのパターンは何かすぐに聞き取れるメカニズムが脳内にあるか、そういう話があります。 /// That is such thing. There is the pattern, and there is such a mechanism that the mechanism which can hear it immediately in the brain." The video player controls at the bottom show the video is at 00:35 of a 01:16:58 video, with a progress bar at 272/291.

# デモ

## • 音情報処理 (2016.10.4)

音声の知覚

- ▶ 明瞭度と了解度
- ▶ 知覚単位と文脈
- ▶ カテゴリー知覚

そういうふう。パターンがあってでそのパターンは何かすぐに聞き取れるメカニズムが脳内にあるか、そういう話があります。 // That is such thing. There is the pattern, and there is such a mechanism that the mechanism which can hear it immediately in the brain.

# 背景

- 高等教育グローバル化の進展による日本語講義・英語講義の混在
  - 日本語で専門知識を学べる
  - 英語で専門知識を学べる
- NAIST情報科学領域の現状
  - **およそ半数**が英語開講科目
    - 全学では7割程度の科目で英語開講あり（日英別クラスとする場合含む）

※予稿での記載が不正確であったため訂正

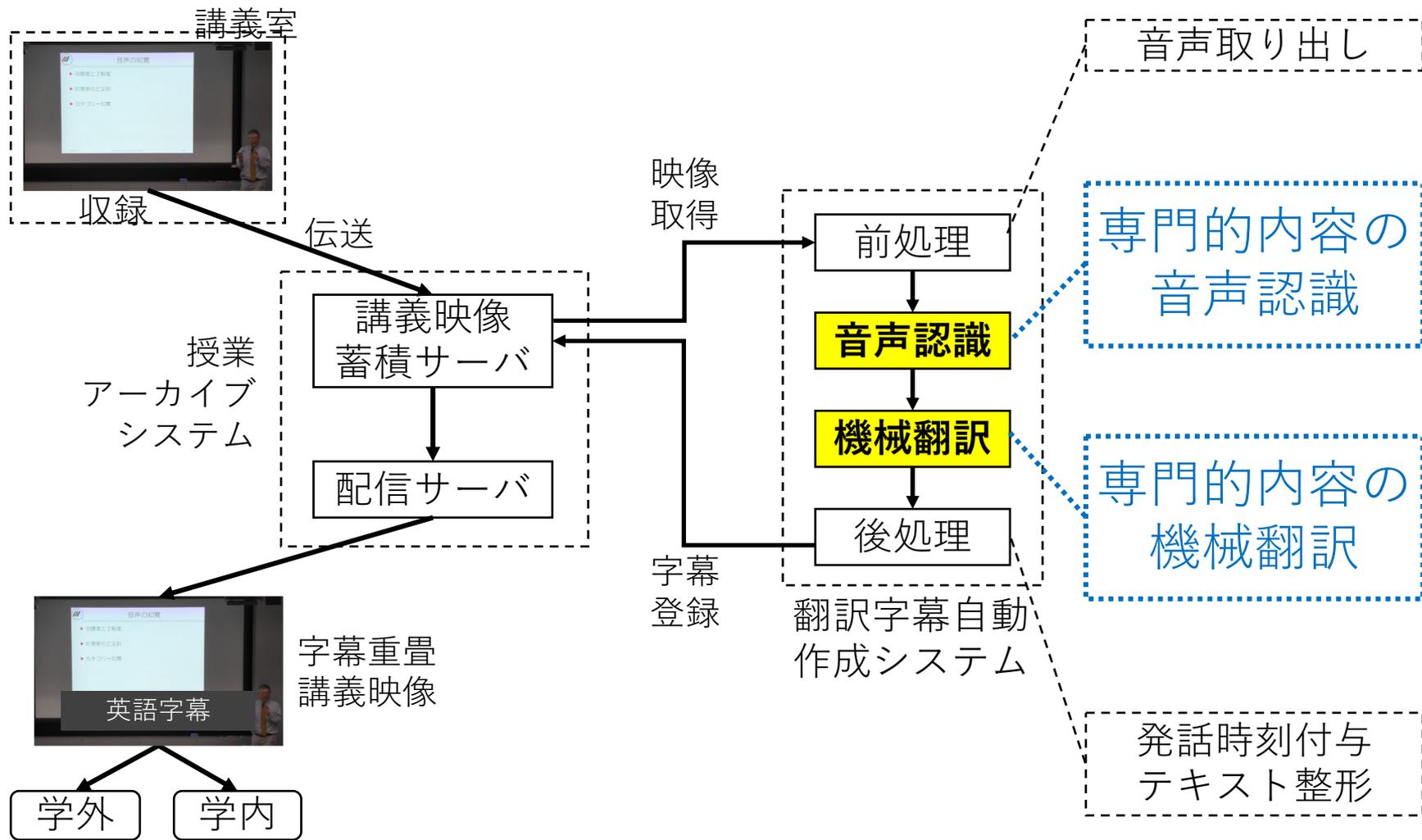
# NAIST授業アーカイブシステム

- 2004年度から収録・学内公開開始，  
2008年度から学外公開開始（一部）
- 翻訳字幕自動作成プロジェクト：  
「日本語講義に英語字幕を付与」
  - 2016年度開始
  - データ整備，字幕重畳表示機能の追加
  - 2018年度試作システム作成

# 本プロジェクトの目的

- 日本語開講科目の学習補助
  - 授業アーカイブの更なる活用促進
- 実用に資する音声・言語処理技術の研究加速
  - 実用における種々の問題
  - 蓄積された音声言語資源の活用

# システム構成



# コーパスの収集

- 2016年度を中心に35講義 約46.5時間
  - 情報科学分野の専門科目 6科目
- 日本語音声書き起こし
  - 時刻情報の付与（約半数）
  - Disfluency（フィラー・言い誤り）タグ
- 英語への翻訳（Disfluency削除後）
  - おおよそ発話単位（句点区切り）
  - 翻訳者による分割・併合も許容

# コーパスの例（書き起こし）

1922 4	2762 4	(F でー)(F えーっとー)この授業では(F あの)音声っていうのは何かということで、(P)(F えー)授業を順番にしていきたいと思います。
2957 4	3782 4	(F え)全体で八回ですけども、(F えーっとー)(F ま){  おんせい かん}音声に関するいろんな(F えーっと){  ぎじゅつ}構成技術 の紹介をしていきます
3858 7	6672 3	(F でー)(F えーっと)(F まあ)音声、人のコミュニケーション で{  い}意図を伝える最も重要な手段ということで、(P)(F えーっとー)(F ま)(F あの)テキストで(F あのー){  い}意図を伝 えるっていうのもありますけども、テキスト以外に(F まあ) 感情とか、それから(F えーっとー)(F お)(F い)(F まあ)強調と かそれから男性女性とか、(F ま)そういった(F そのー)いろい ろな(F あ)音声に含まれている(F あ)全体の情報に対して、(F ま)それを処理して(F えっと)どうやって{  つたえ}伝わって行 くのかっていうことを(F まあ)研究すると(P)いうものです。

# コーパスの例 (翻訳)

今日の内容はですね、音声対話、音声対話についての講義を行います。	Today's content is speech dialogue, I am giving a lecture about speech dialogue.
今日の内容です。	This is today's content.
まず初めにですね、音声対話システムというものがどういうものかということですね、簡単に説明します。	First of all, I will explain simply what the speech dialogue system is like.
音声対話システムっていろいろな要素技術が使われてて、全部を説明するのはちょっと不可能なんですね。	Since various underlining technologies are used for speech dialogue, it is quite impossible to explain everything.
なので、今回はその中でもですね、一番核となる部分ですね、対話制御の部分にちょっと注目して講義を進めたいと思います。	So, this time in this lecture I will focus on the core part, on dialogue control.

# 音声認識部

言語モデルのみの  
ドメイン適応がしやすい

- Kaldi [Povey+ 2011]
  - 音響モデル：DNN-HMM [Moriya+ 2015]
    - 日本語話し言葉コーパス240時間で学習
  - 言語モデル：単語 3-gram
    - 日本語話言葉コーパス 20万文
    - ATRコーパス 57万文
    - Web収集コーパス 10億文
    - 過去3年の授業アーカイブ音声認識結果  
(パープレキシティで選択) 52万文
  - 語彙サイズ 26.4万

# 機械翻訳部

- 注視型ニューラル機械翻訳 [Luong+ 2015]
  - LSTMの層数は1 (データ量の問題)
  - ASPEC [Nakazawa+ 2016] で事前学習
  - 授業アーカイブコーパスで追加学習
  - SentencePiece [Kudo+ 2018] でサブワード化
    - 語彙は日英共有, サイズは16,000
- 翻訳時ビーム幅: 10

# 予備実験：設定

- テストセット：3講義
- 開発セット：講義コーパスの一部
- **音声認識評価：単語誤り率**
  - ルールベースの表記揺れ吸収
  - 言語モデルclosedの結果も比較
- **機械翻訳評価：BLEU**
  - 書き起こしに対する機械翻訳
  - テストセット内の計500発話分

# 予備実験：結果（音声認識）

- 単語誤り率は10%台
  - Disfluencyの影響が無視できない
  - 言語モデルのカバレッジ不足

講義名	単語誤り率(%) ↓	単語誤り率 (%) ↓ (言語モデルclosed)
ロボティクス	12.48	8.25
音情報処理	12.56	6.65
ソフトウェア工学	17.76	13.31

表：各講義1回分の音声認識結果の単語誤り率

# 予備実験：結果（機械翻訳）

- 書き起こし翻訳でもBLEUで10%台
  - 訳が短くなる傾向
    - ビーム探索によりある程度BLEUは向上

講義名	BLEU (%) ↑	Brevity Penalty (%)
ロボティクス	15.9	90.8
音情報処理	21.1	89.7
ソフトウェア工学	12.8	87.2

表：各講義の書き起こしに対する機械翻訳結果のBLEU（合計500発話）

# 課題

- 認識・翻訳通しで適用すると...
- 個々のモジュールの基本性能
  - 固有表現・専門用語のカバレッジ不足
  - 講義の発話スタイルへの適応が不十分
- Disfluencyの影響
  - 音声認識の乱れが生じやすい
  - 訳抜けや不必要な句の繰り返しを誘発

# 今後の展望

- ドメイン適応
  - 他分野の講義への拡張
- データ拡張
- 英語講義の英日翻訳
  - 非母語話者音声認識
  - 文法誤り訂正

# まとめ

- 日本語授業の音声翻訳システム試作
  - 講義コーパスの収集 約46.5時間分
  - Kaldiベースの音声認識
  - 注視型NMTによる機械翻訳
  - モジュール毎の予備実験と評価
- 単純なカスケード処理では不十分
  - 音声認識誤りの影響
  - Disfluencyの影響