

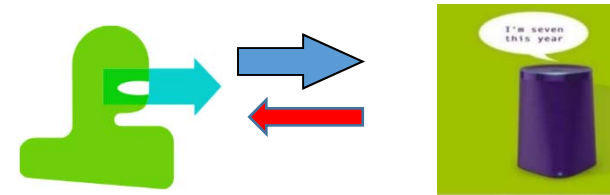
アンドロイドERICAによる 人間レベルの音声対話

河原達也
(京都大学)

現状の音声対話システム

- 情報検索・機器操作システムとのインタフェース

- タスクに沿った [概念的制約]
- 単純な文を [言語的制約]
- 明瞭に発声 [音響的制約]
- 応答を待つ [受動的対話]

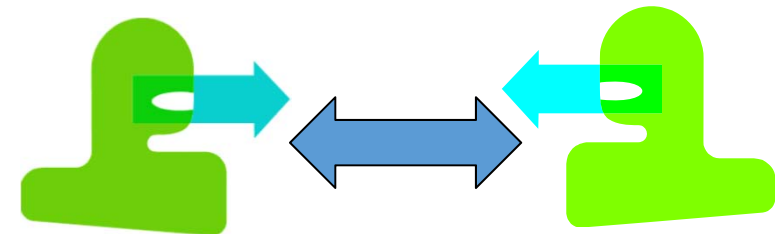


Half duplex




- 人間どうしの対話

- タスクが明確でない
- 1ターンで多数の発話／相槌
- **考えながらやりとり**
- 対話を通じてお互いの考えが明確になる



Full duplex

タスクを遂行するための対話

- 明確なタスク
 - 機器操作
 - 手配
 - 検索

瞬時にできる(すべき)もの
- 評価尺度
 - 意図に沿った応答
 - 客観的に正解が定義可能
 - 迅速に
 - 瞬時が望ましい
- 対話はあくまでも手段
 - 会話ロボット向けでない [今井18]

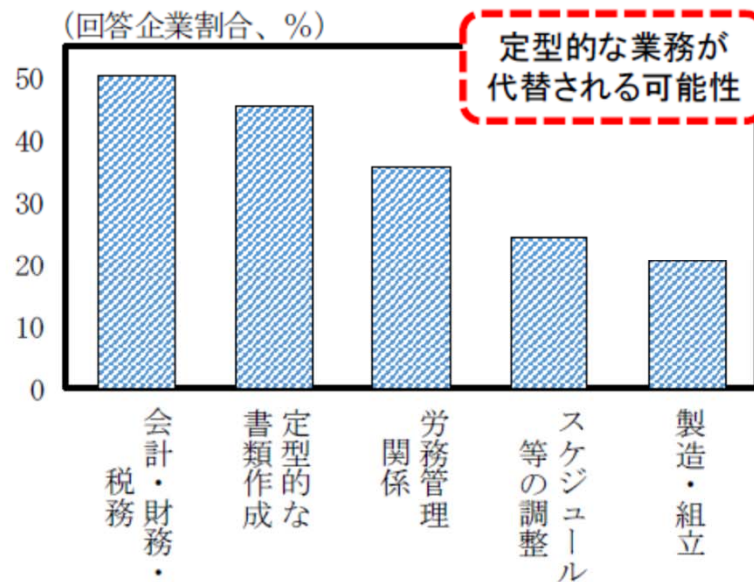
対話自体が目的となるタスク

- ゴールが明確でないが、
雑談(時間つぶし・社交目的)ではない
(cf.) 目的をもった面会
- 評価尺度
 - 長く話す
 - エンゲージメント(対話感)
 - (客観的に定義できない)タスク

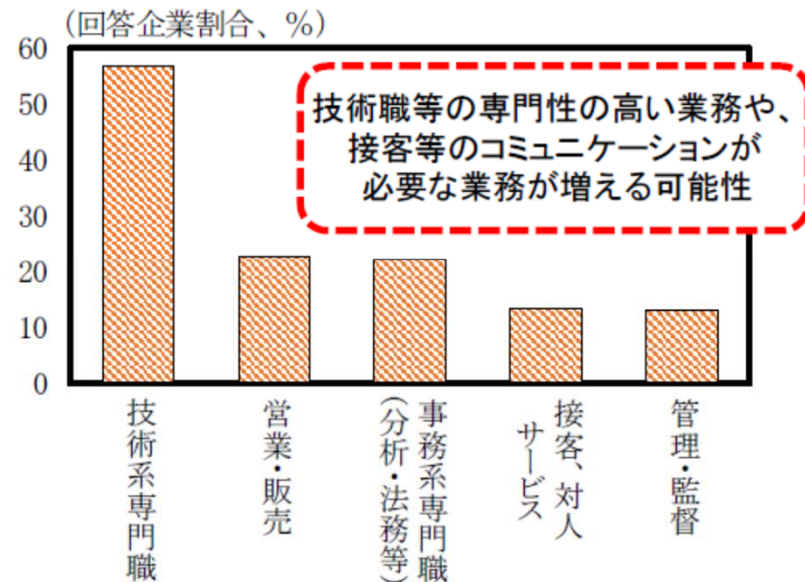
2018経済財政白書

今後AI等の進展により、定型的な業務が代替される一方、専門性の高い業務や接客・対人サービス等のコミュニケーション能力が必要な業務(の人材需要)が増える

(1) 企業がAI等に代替を考えている業務



(2) 企業がAI等により増えると考える仕事



AIに容易に代替されないと考えられるタスクが究極のAIの目標

Android ERICA



JST ERATO 石黒共生ヒューマンロボット インタラクションプロジェクト (2014-2020)

- **目標:** 人間と同様にインタラクションできる自律型アンドロイド
 - 表情・視線・頷き
 - 音声対話
- **究極的目標:** Total Turing Test
 - 人間と同様の対話感
 - 遠隔操作のアンドロイドと区別できないレベル
- **科学:**
 - 自然な対話において何が不可欠で、現状何が不足しているのか
- **工学的応用:**
 - 対人コミュニケーションタスク
 - 人間のコミュニケーションスキルの訓練

人間レベルの音声対話

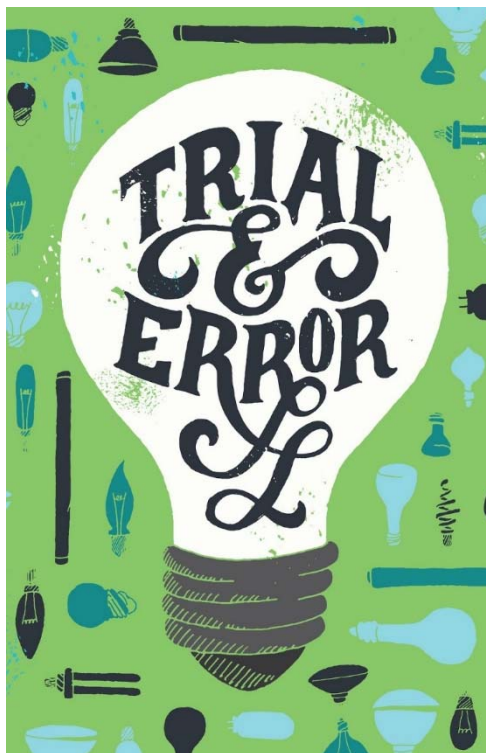
- × 情報検索・サービス → スマートフォン
- × 物体の移動など → 従来のロボット
- × 雑談 → チャットボット
 - 基本的に一問一答



- 長い深いやりとり
- 人間らしい存在感
- 非言語情報を含む対話感

対話のタスク

当初: 研究室案内・受付・秘書システム



- 対話感なし
- ユーザ発話低調

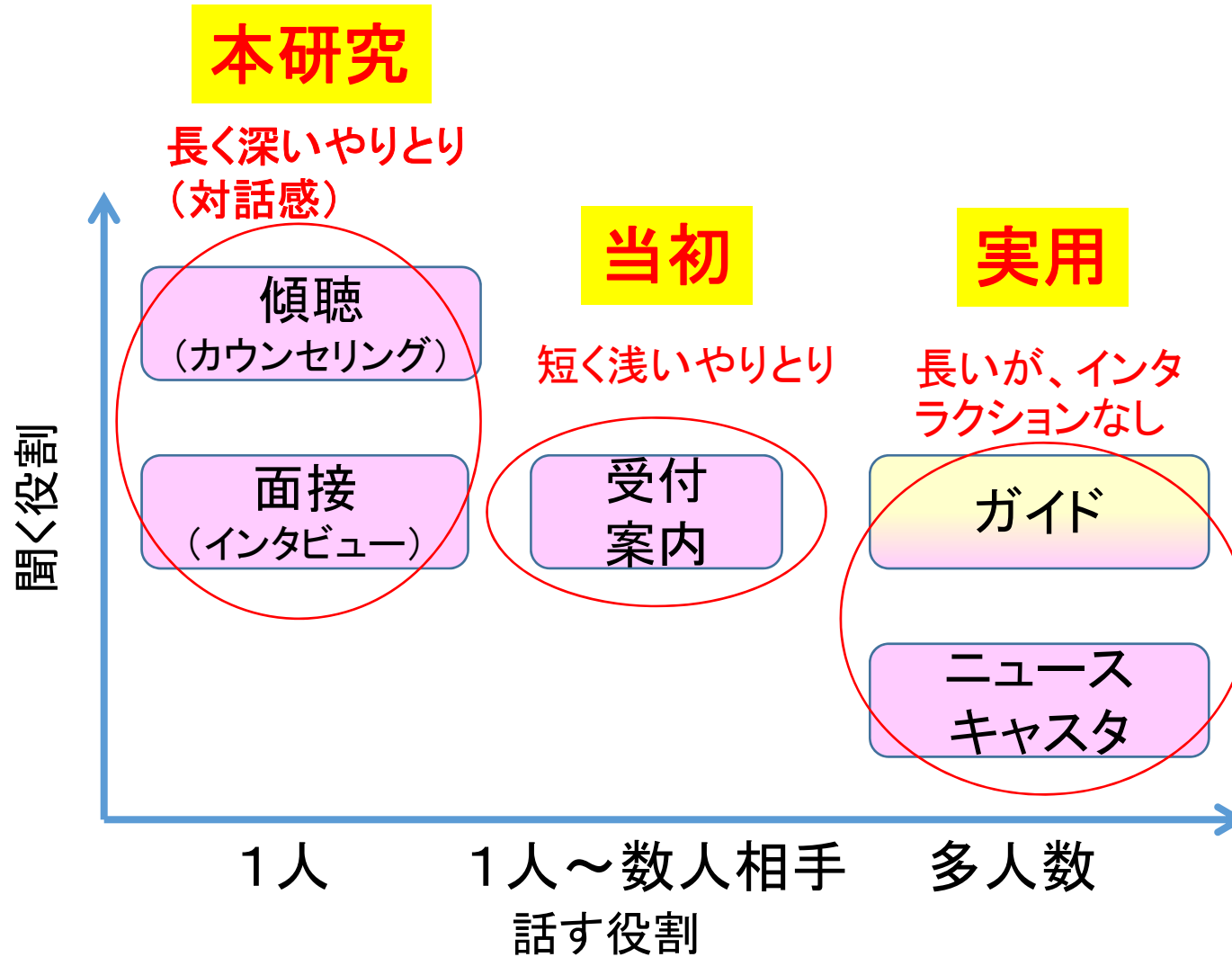
ゴールが明確でない対話システムの問題

- システム応答が、**つまらない**(無難) OR **見当違い**(無謀)
- 分別のある大人(大学生)は、ロボットと対話したがる

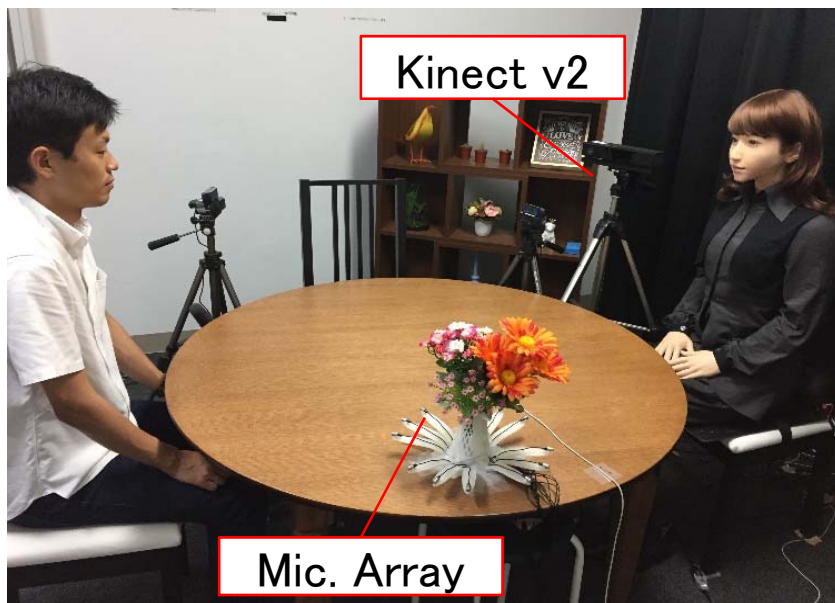


- システムに明確な社会的タスク
 - 単なる雑談でない
- ユーザにリアリスティックな設定
 - 真剣に対話に臨ませる

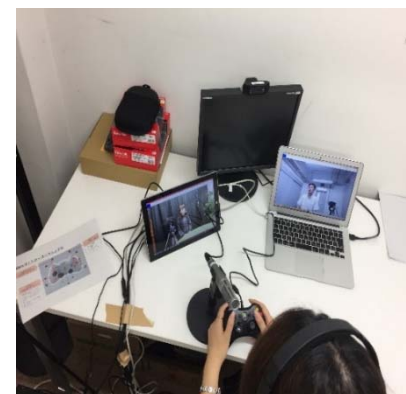
コミュニケーションロボットによる対話タスク



WOZによる対話データの収集



応答
制御



Task 1: 傾聴

- 高齢者の話を聞く [下岡17]
 - 印象に残った旅行、最近行っていること
 - 的確なフィードバックを行うことで、円滑な発話を促進
 - カウンセリング [DeVault14, 河原15] とも類似



大半は、遠隔操作に
気づいていない

Task 2: 就職面接(練習)

- ERICAが面接官役
 - 志望動機やスキルなどについて質問
 - 応答に応じた追加・掘下げ質問
 - インタビュー [小堀16, 長澤17] と類似
- ユーザ(学生)はアピールする必要 → 実際のシミュレーション



かなり緊張

人間らしい存在感
が重要

Task 3: お見合い(練習)

- ERICAが女性参加者役
 - 趣味や好きな食べ物などの話題について、ユーザに質問したり、ユーザの質問に答える
 - 対話に応じたフィードバック
- ユーザ(男子学生)はアピールするだけでなく話を聞く必要
→ 実際のシミュレーション



リラックスしているが、
それなりに真剣

人間らしい存在感
が重要

Face-to-Faceコミュニケーション

- 傾聴
- 面接
- 面談
- お見合い

Face-to-Faceコミュニケーション が必要不可欠な場合

- 深刻な相談
 - カウンセリング
 - 面談
- 人物の評価
 - 入学試験の面接
 - 就職面接
 - お見合い

コミュニケーションスキル

- **話す**（聞いてもらう） → **ガイド**
 - 一方的に話すのではなく、相手に興味をもって聞いてもらう
- **聞く**（話してもらう） → **傾聴・カウンセリング**
 - 的確にフィードバックすることで、相手に話し続けさせる
- **尋ねる** → **面接・面談**
 - 相手から情報を引き出す
- **答える** → **相談**
 - 答えるためのDB・KBが必要
 - 答えるには尋ねる必要
- **実際には**
 - 上記の組合せ → **お見合い**
 - **ノンバーバル**（デリバリ・視線）も重要
- 本研究では各々に焦点が当たるタスクを構成的に設計・実装

3つのタスクの比較

	傾聴	就職面接	お見合い
システムの役割	聞く	尋ねる	すべて
対話の主導権	ユーザ	システム	両方(混合)
発話の大半	ユーザ	ユーザ	両方
相槌の大半	システム	システム	両方
発話権交替	あまりない	明確	複雑

3つのタスクの比較

	傾聴	面接	お見合い
収録対話数	19	30	33
ユーザ発話の割合%	64%	53%	49%
相槌生起の割合%	38%	19%	19%
ターン切替の割合%	19%	30%	37%
ターン切替時間	-34msec	365msec	120msec

対話の構成要素

相槌の生成

- 発話を促進
 - 聞いているというフィードバック
 - “はい”, “うん”
- 感情を表出
 - 驚き・興味・共感
 - “あー”, “へー”
- 対話のリズム・同調性を形成

相槌生成の要素

- タイミング (when) ← 多くの先行研究
 - 発話の終了(区切り)時
 - 発話の終了前に予測する必要
- 形態 (what)
 - 韻律と言語素性を用いた機械学習
- 韻律 (how)
 - 先行発話の韻律にあわせて調整

TTSで
専用の
エントリ

従来システムは同じパターンの繰返しのため単調



相槌の種類と頻度

形態	発話末における頻度
「うん」	12% (10%)
「うんうん」	7% (9%)
「うんうんうん」	13% (19%)
感情表出系「あー」	8% (14%)
なし	60% (47%)

- 約40%の発話末で出現
- 形態も多様

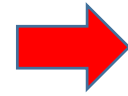
相槌の追加アノテーション

- 相槌の生成と形態は任意性が高い
- コーパスに出現したものが正解とは限らない

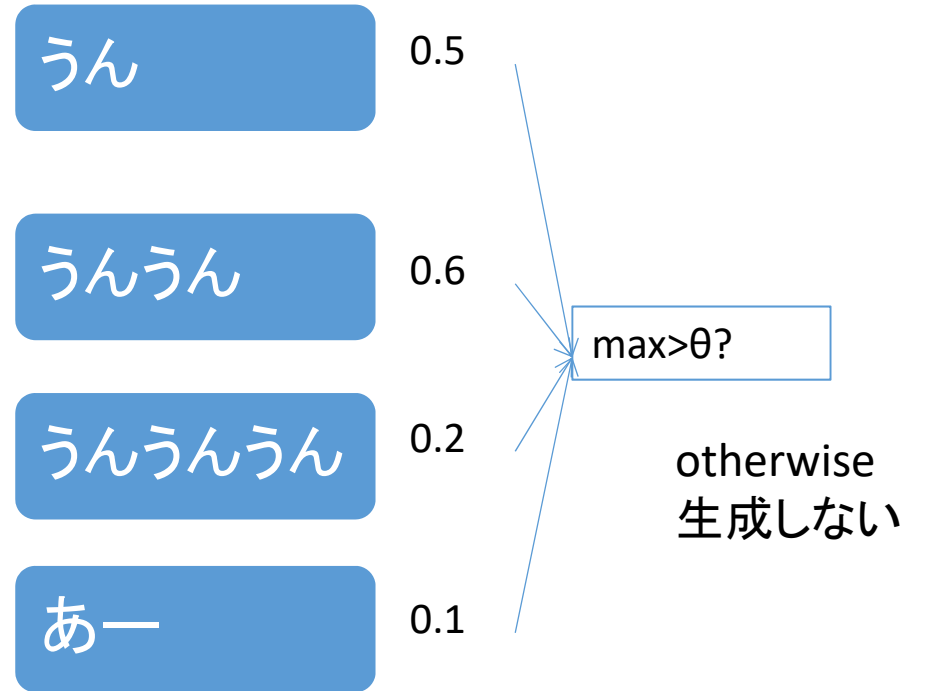
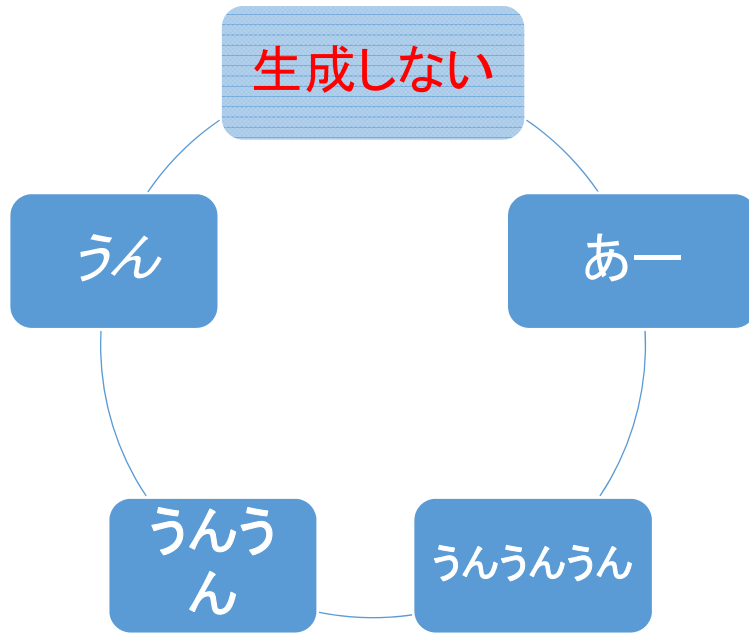


- 妥当なものを追加アノテーション
 - 3名の被験者が一致した場合に採用
 - 評価における正解とする

選択問題



妥当性判定



相槌の予測精度の評価

カテゴリ	Recall	Precision	F-measure
「うん」	0.311	0.657	0.422
「うんうん」	0.382	0.820	0.521
「うんうんうん」	0.672	0.333	0.454
感情表出系「あー」	0.467	0.342	0.405
なし	0.775	0.769	0.772
平均	0.643	0.643	0.643

- 応答系の相槌の適合率は高い
- 生成しない場合の予測精度も高い

相槌の主観評価 (-3~+3)



	ランダム	提案予測 モデル	人間のカ ウンセラ
相槌は 自然 ?	-0.42	1.04	0.79
対話の テンポ はよい?	0.25	1.29	1.00
親身に聞いているか?	0.33	1.25	0.96
理解 してくれている?	-0.13	1.17	0.79
関心を持って聞いている?	0.21	1.21	1.04
共感 してくれている?	0.13	1.04	0.46
このカウンセラと話したい	-0.33	0.96	0.29

赤字: ランダムに比べて統計的有意 ($p < 0.01$)

予測モデルは人間のカウンセラと同等の評価 (TTT)
ただし、韻律の問題はあり

焦点語に基づく聞き返し

- 相槌や語彙的応答のみでは、対話の維持困難
- オープンドメインにおいては、的確な質問の生成も困難



- ユーザ発話から**焦点語**の抽出
 - 音声認識結果の信頼度高い
 - 比較的長い名詞
- 繰返し
 - (例)「この前インドに行きました」→「インドですか」
- 質問
 - (例)「そこでカレーを食べました」→「どんなカレーですか」

焦点語に関する 掘り下げ質問／繰り返し応答

- 「昨日の晩はカレーを食べました。」

単語連鎖確率を計算

○どんなカレー △誰のカレー
△いつのカレー △どこのカレー

5W1H

↓
「どんなカレーですか？」(掘り下げ質問)

→実際にはユーザが沈黙したときに生成

- 「昨日の晩はうなぎを食べました。」

単語連鎖確率を計算

△どんなうなぎ ×誰のうなぎ
×いつのうなぎ △どこのうなぎ

5W1H

↓
「うなぎですか」(繰り返し応答)

評価応答の生成

- 各名詞に付与された感情極性値を集計
- 以下のいずれかの値が一定以上になれば応答

	肯定的	否定的
客観的（事実）	素敵ですね	大変ですね
主観的（意見）	いいですね	残念ですね

「海に行きました」→「素敵ですね」

「でも疲れました」→「残念ですね」

- ある程度対話が進行してから生成

その他の構成要素

- 質問応答・挨拶
 - 想定されるもの
- 語彙的応答
 - 上記のいずれでも対応できない場合
(例)「そうですか」「なるほど」
- 状態遷移モデル
 - 質問のリスト／フロー
 - 大局的な対話の流れを記述
 - ユーザが沈黙してしまった場合

ターンテイキング

- 既存の対話システム：発話できる区間を指定
 - スマートフォン：発話時にクリック (push-to-talk)
 - スマートスピーカ：マジックワード “Alexa”, “OK Google”
 - 一部のロボット：LED点滅時に発話



- 人間どうしの対話
 - 発話交代の時間：0(傾聴)～400msec(面接)



- 人間らしいシステム
 - 相槌やフィラーの生成と統合

傾聴システムの構成

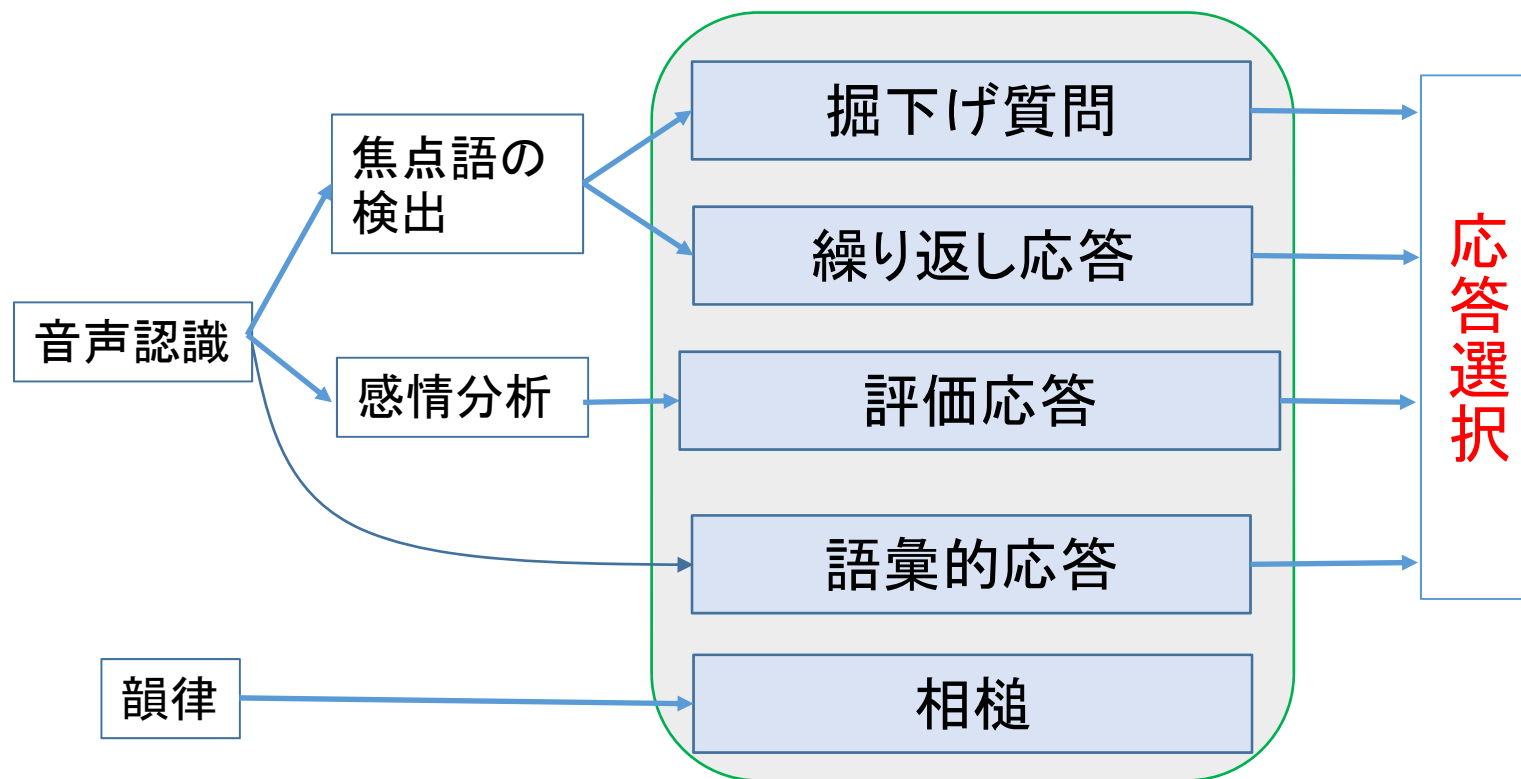
傾聴システムの仕様

- 聞くことに関してオープン（理解しているわけではない）
 - 「旅行」「食べた物」「健康法」などの話題は与えるが、システムはどんな話題でも対応可能
- システムから質問は（原則）行わない
 - ユーザが質問を待つモードになるのを防ぐ
- 高齢者による5分の対話を目標

傾聴システムの主要要素

- **自然な相槌** → “話を聞いてもらっている”感覚
 - 多様な相槌を選択
 - 「うん」「うんうん」「うんうんうん」「あー」
- **聞き返し** → “話を理解してもらっている”感覚
 - 焦点語の検出 「* * ですか」
 - 掘下げ質問の生成 「どんな* * ですか」
- **評価応答** → “話に共感してもらっている”感覚
 - 「素敵ですね」「大変ですね」

傾聴対話システムの構成



応答選択

- 複数の応答候補が生成
- 正解があるわけでない(コーパスも正解とは限らない)

「この前の日曜に高校の同窓会に行きました」

語彙的応答	「そうですか」	○
評価応答	「素敵ですね」	○
繰り返し応答	「同窓会ですか？」	○
掘下げ質問	「どの同窓会ですか？」	×



1つを選択するのではなく、個々の妥当性を判定

応答選択

- コーパスに出現したもの以外にも可能な候補は多数



- 対話文脈を与えて妥当性をアノテーション

	コーパス出現	→ 妥当な割合
語彙的応答	45%	90%
評価応答	21%	60%
繰り返し応答	22%	64%
掘下げ質問	11%	28%

- 語彙的応答はたいてい可能
- 評価応答と繰り返し応答も過半数で可能

応答生成・選択の評価

	Recall	Precision	F-measure
語彙的応答	99%	91%	0.95
評価応答	51%	73%	0.60
繰り返し応答	68%	80%	0.74
掘下げ質問	46%	41%	0.43
重み付き平均	70%	73%	0.71

- チャンスレートより有意に向上
- ただし、かなりの不適切な掘下げ質問が生成
→ 掘下げ質問は間があいたときのみ出力

システムのデモ

- うまくいく対話
 - 適度に聞き返し、最後に評価応答
- 何とか乗り切る対話
 - ほとんど相槌のみ
- うまくいかないユーザ
 - 話すことがなくなる
- リスponsがもう少し早いとよい
 - 発話終了前に相槌を予測・生成する必要
- [傾聴デモビデオ](#)

実際のシステムの評価

京都大学カウンセリングルーム：杉原保史教授
『プロカウンセラーの共感の技術』の著者

5分の傾聴対話を体験（一応成功）後、
「子どもと話しているようだ」

ある話題について5分間話せるか？

- 犬・猫



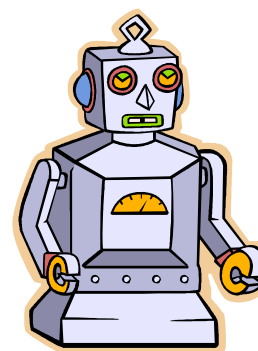
- 幼児



- 子供(小・中学生)



- ロボット



- ERICA(23歳)



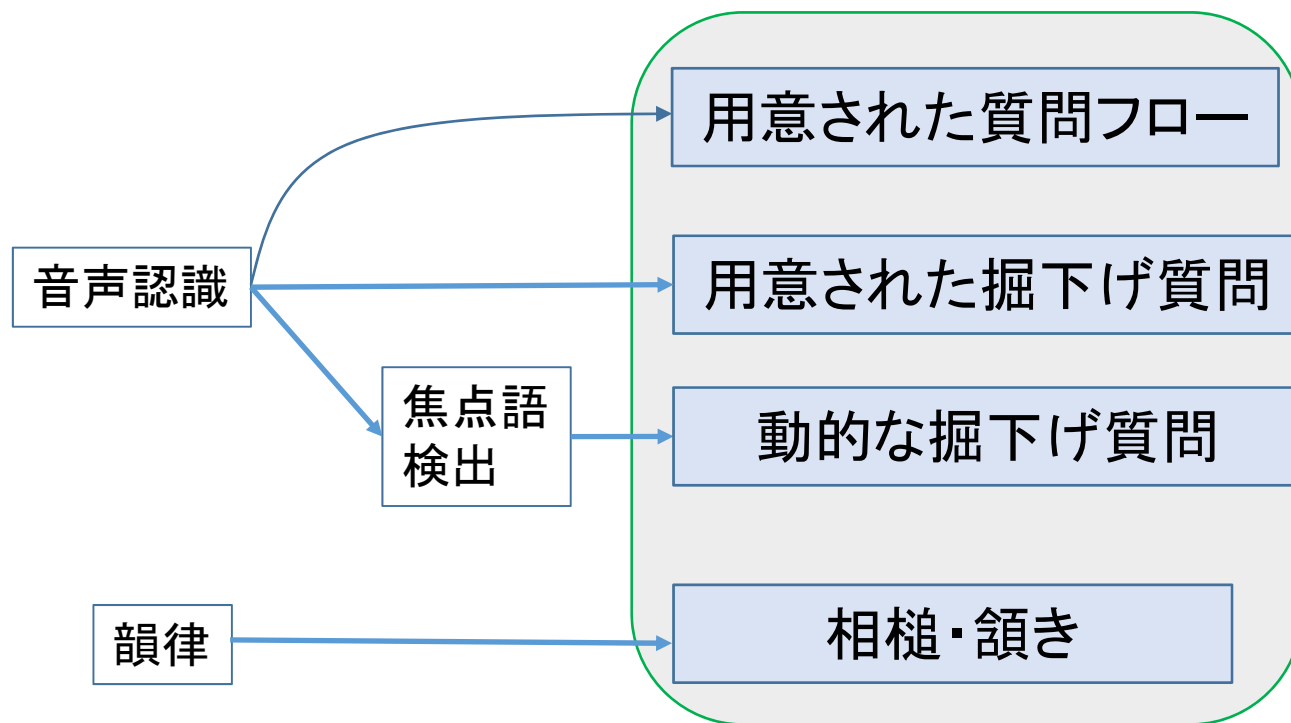
現在はこのレベル(TTT)

就職面接システムの構成

就職面接システムの仕様

- ユーザは志望する企業・業種を想定して対話するが、システムはどんな業種・企業でも対応可能
- 基本フロー
 - 志望動機
 - 学生時代に頑張ったこと
 - その他(スキルなど)
- 応答に応じた掘下げ質問
 - 用意された候補からの選択
 - (例)「当社でないといけない理由はあるのでしょうか」
 - ユーザ発話中のキーワードの掘下げ
 - (例)「深層学習についても勉強してきました」
 - 「では、深層学習について説明して下さい」

就職面接システムの構成



IROS-2018 workshopでのデモ



システムのデモ

- ほとんど破綻しない
 - ユーザは明瞭に発話
 - システム主導の対話
- 一見おかしな掘下げも哲学的にとらえられる
 - 「研究」や「チーム」に対する質問
- [就職面接デモビデオ\(日本語\)](#)
- [Job interview demonstration video \(English\)](#)

実際のシステムの評価

- 京都大学キャリアサポートルーム: 松尾准教授
「質問が制御できないと指導目的には難しいが、
自主練習には使えそう」

今後の課題

- お見合いシステム
- 本格的な理解
 - 現在は内容語の抽出のみ
 - 発話行為タグ
 - オープンドメインにおける「理解」の定義
- メンタルモデル
- 視線や表情などの利用

For Demo Video

Search “[2018 ERICA @ kyoto-u](#)”