第2回：ヒトはどのように声を出し、聴くのか

音声処理と認識

担当：小野原彩香

形式：オンライン同期（Zoom）

Zoom録画を後日公開します。課題はメール/Google Driveでも提出可。

今日のアジェンダ

1. 前回の振り返りクイズ（Zoomチャット）
2. 発話器官のしくみ
3. 調音音声学の基礎
4. 音声の音響的特性（ソース-フィルタモデル・フォルマント）
5. 聴覚系のしくみ
6. 人間の音声知覚の特性
7. 人間と機械の比較
8. グループディスカッション
9. まとめ・小課題

事前学習の確認クイズ

Q1：「マクガーク効果」とはどのような現象か？（Zoomチャットに書き込む）

A) 騒音の中でも特定の人の声を聞き分けられる現象
B) 映像の口の動きが聴こえる音の知覚に影響を与える現象
C) 自分の録音した声が普段の声と違って聞こえる現象
D) 耳介の形が音の方向知覚に影響を与える現象

Q2：自分の録音した声を聴いて、どう感じましたか？（自由記述）

発話は複数の器官の協調動作

> 肺 → 気管 → 喉頭（声帯）→ 咽頭。咽頭から口腔側・鼻腔側へ分岐し、軟口蓋が鼻腔との接続を調節する。

器官	発話における役割
肺・呼吸筋	呼気を生み、音を作るためのエネルギーを供給する
声帯	有声音では振動し、音源となる気流の変動を作る
声道	声の通り道。形状によって共鳴周波数（フォルマント）が変わる
舌	口腔内の形状を変え、子音・母音の調音に関わる
唇	開閉・突き出しなどで音を変える。両唇音 [p][b][m] の形成
鼻腔	軟口蓋が下がると声道に接続。鼻音 [m][n][ŋ] では口腔を閉じ、気流を鼻へ通す

声のしくみ：ソース-フィルタモデル

> **有声音の基本モデル**：声帯振動（音源）→ 声道（フィルタ）→ 放射される音

ソース（音源）

- 肺からの呼気がエネルギーを供給し、声帯振動が周期的な音源を作る
- 基本周波数 F_0 は、1秒あたりの繰り返しの回数。音の高さの主な手がかり
- F_0 = 100 Hz なら1周期は10 ms、200 Hz なら5 ms
- 無声摩擦音では乱流、閉鎖音では開放時の破裂などが音源になる

フィルタ（声道）

- 声道の長さや形が、共鳴周波数（**フォルマント：F1・F2 など**）に関わる
- **F0（音源の基本周波数）と F1・F2（声道の共鳴周波数）を区別する**

有声音と無声音

有声音（Voiced）：声帯振動を伴う音

- 通常の母音、[m]・[n]・[z] など
- 日本語の母音も、発話環境によって無声化することがある

無声音（Voiceless）：声帯振動を伴わない音

- [s]・[p]・[t]・[k] など
- 「サ」は、無声子音 [s] と通常は有声の母音 [a] からなる。**どの音の部分かを区別する**

> **体験実験**：母音を付けずに [sː] と [zː] を伸ばし、喉の振動を比べよう。[sː] は無声、[zː] は有声。ː は音を長く伸ばす記号。

子音の分類：調音様式と調音部位

調音様式（どのように音を作るか）

調音様式	説明	日本語の代表例
閉鎖音（破裂音）	口腔内を閉鎖し、開放する	「カ・タ・パ」の子音 [k][t][p]
摩擦音	狭めた部分で乱流を作る	「サ・シ・フ」の子音 [s][ɕ][ɸ]
破擦音	閉鎖に続いて摩擦を生じさせる	「チ・ツ」の子音 [tɕ][ts]
鼻音	口腔を閉鎖し、気流を鼻へ通す	「マ・ナ」の子音 [m][n]
弾き音	舌先が歯茎に瞬間的に触れる	ラ行の代表的な子音 [ɾ]
接近音（半母音）	強い摩擦を生じさせずに狭める	ヤ行・ワ行の子音

調音部位（どこで音を作るか）

両唇：[p][b][m] ／ 歯茎：[t][d][n][s][z] ／ 歯茎硬口蓋：[ɕ][tɕ]
硬口蓋：[j] ／ 軟口蓋：[k][g][ŋ]　※摩擦音では気流を完全には閉鎖しない。

母音を聞き分ける手がかり：フォルマント

- **F1（第1フォルマント）**：おおむね舌が低いほど高くなる（「ア」など）
- **F2（第2フォルマント）**：おおむね前舌の母音ほど高くなる（「イ」など）
- 声道の長さ・唇の形なども影響する。音程の主な手がかり **F0 とは別の量**

母音の違いを考えるための模式表（同じ話者の中での相対的な傾向）

母音	F1 の傾向	F2 の傾向
ア	イより高い	イより低い
イ	低い	高い
ウ	アより低い	イより低い
エ	イより高く、アより低い	イより低く、オより高い
オ	ウより高く、アより低い	イ・エより低い

値や分布は話者・発話条件で変わる。日本語の「ウ」は、典型的な強い円唇の [u] と同一ではない。

耳はどのように音を処理するか

> 外耳 → 鼓膜・中耳の耳小骨 → 内耳の蝸牛 → 聴神経 → 脳

区域	主な器官	聴覚における役割
外耳	耳介・外耳道	音波を集め鼓膜へ伝える。耳介の形は方向知覚の手がかりになる
中耳	耳小骨（3個）	鼓膜からの振動を、内耳の液体へ効率よく伝える
内耳の聴覚器官	蝸牛	周波数に応じて異なる場所が応答し、振動を神経系の信号へ変換する

鼓膜は外耳と中耳の境界にある。内耳には、平衡感覚に関わる**前庭・半規管**もある。

蝸牛：カタツムリ型の周波数分析器

> 蝸牛の入口に近い**基底部は高い周波数**、奥の**頂部は低い周波数**に応答しやすい。

蝸牛の重要な特性

1. **周波数と場所の対応（Tonotopy）**：周波数に応じ、基底膜の異なる場所が振動する
2. **機械的な振動から神経系の信号へ**：有毛細胞の電位が変化し、聴神経へ情報が伝わる
3. **非線形な応答**：入力音の強さと応答の大きさは単純には比例しない

メル尺度は、周波数と知覚される**音の高さ**の関係を表す心理音響的な尺度。

> 周波数ごとに分ける働きはフィルタバンクなどと比較できる。ただし、蝸牛がフーリエ変換と同じ計算をしているという意味ではない。

音声は「目」でも聴く ― マクガーク効果

1976年、McGurk & MacDonald が報告した現象

実験の例

- 「ガ（ga）」を発音する口の映像に、「バ（ba）」の音声を重ねる
- 音声とは異なる「ダ（da）」などに聞こえることがある
- 目を閉じて音だけを聴いた場合と比べる。効果の出方は人・言語・刺激によって異なる

この現象が示すこと

- 口の動きなどの視覚情報が、音声の聞こえ方に影響する
- 音声だけを入力するASRは視覚情報を使わないが、**音声と映像を入力する認識モデル**もある

騒音の中で声を選別する — カクテルパーティ効果

大勢が話す環境でも、特定の人の声に注意を向けて聞き取れる現象

Cherry（1953）：左右の耳に異なるメッセージを提示する両耳分離聴実験

手がかり・処理	人間	機械
聞き取る話者の選択	声の方向・声質・注意などを使う	マイクアレイ、対象話者の声の特徴、口の映像などを利用できる
事前知識・文脈	話者や話題の知識を使って解釈する	言語モデルや事前に与えた語句などを利用できる
雑音・重なった声	注意や聴覚的な分離が働くが、条件によって聞き取りは低下する	音声強調・話者分離を用いるが、雑音や重なりの影響を受ける

技術例：ビームフォーミング・Wienerフィルタ・深層学習による話者分離／対象話者の抽出

人間も機械も、入力条件と利用できる手がかりによって成績が変わる。

音声知覚の「不思議」な現象たち

カテゴリカル知覚（Categorical Perception）

- 音響特性を連続的に変えても、分類の割合や聞き分けやすさがカテゴリ境界付近で大きく変わることがある。同じカテゴリ内の違いも知覚される

音素修復効果（Phonemic Restoration）

- 文中の音を雑音で置き換えても、文脈などを手がかりに本来の音が続いているように知覚されることがある

腹話術効果（Ventriloquism Effect）

- 音が聞こえる位置が、同時に示された視覚情報の位置に引き寄せられる

プライミング効果

- 先に提示された情報が、後の刺激の認識や反応に影響する

人間の音声知覚と機械の音声処理

ASRは主に音声を文字列に変換する技術。 感情・意図の推定は関連する別の課題。

比較軸	人間の音声知覚	機械（ASRと関連する技術）
周波数分析	蝸牛の周波数に応じた応答など	短時間フーリエ変換・メルフィルタバンク、波形から学習した特徴など
処理の時間	聞きながら逐次処理する	録音全体を処理する方式と、逐次処理する方式がある
雑音・声の重なり	注意などを使うが、聴取条件に依存する	モデル・学習データ・入力条件などに依存する
方言・学習	長期の言語経験を基に、新しい話者などに適応する	大規模な事前学習を基に、少量の文字起こし付きデータで適応できる場合がある
視覚・文脈	口の動きや文脈も知覚に影響する	文脈や映像を利用する認識モデルもある
感情・意図の推定	韻律・文脈などから推定し、誤解も起こる	感情認識・意図推定のモデルを使い、文字起こしとは別に評価する

比較するときは、課題・データ・評価条件をそろえる。

グループディスカッション（20分）

実施方法

- 希望するZoom参加者はグループ（3〜4人）に分かれて議論する
- 20分後に各グループが2〜3分で発表する
- 発言が苦手な場合はテキストチャットでの参加でも構わない
- Zoomに参加できない場合は、テーマを選んでメールで意見を送ってもよい

> 発表は「正解を言う」ことが目的ではありません。議論のプロセスと考え方を共有してください。

テーマ割り当て（以下のスライドを参照）

ディスカッション：テーマA

「機械に人間のカクテルパーティ効果を実現させるには？」

- マイクのハードウェア面（マイクアレイ等）で何ができるか？
- 信号処理・機械学習でどんなことができるか？
- 「特定の声を選ぶ」には、事前にどんな情報が必要か？
- 実際の製品（スマートスピーカー・補聴器）はどう対処しているか？

発表でまとめること：考えたアプローチを2〜3つ挙げ、各アプローチの難しい点も述べる

ディスカッション：テーマB

「音声認識のバイアスとはどういう問題か？どう対処すべきか？」

- 学習データの偏りとは何か？（誰がデータを提供したか）
- 声の物理的特性（声の高さ・フォルマント分布）の多様性はどう扱われているか？
- バイアスのある音声認識が社会で使われるとどんな問題が起きうるか？

発表でまとめること：バイアスが生じるメカニズム＋役割別（開発者・利用者・政策）の対策

ディスカッション：テーマC

「マクガーク効果から見えること ― 音声とは何か？」

- 映像と音声の統合は、どのような条件で有利になるか？仮説と検証方法を考える
- マスクや映像の遅延があるとき、聞き取りはどう変わるか？
- 機械で同様の現象を調べるなら、どのような技術と実験が必要か？
- 機械の「誤認識」と人間の「誤知覚」には、どんな共通点・相違点があるか？

発表でまとめること：比較する課題と共通点・相違点を1〜2文でまとめ、最も面白いと思った視点を共有する。

やってみよう！音声知覚の体験実験

授業内では**3つから2つを選択（各5分・計10分）**。残りは任意の復習用。

演習1：有声音・無声音の識別

- 母音を付けずに提示された部分を判定し、Zoomチャットへ記入する
- 例：[zː] 有声／[sː] 無声／[mː] 有声／[ɸː] 無声／[aː] 有声

演習2：F0 とフォルマントを区別する

- できるだけ同じ音程を保って「アー」から「イー」へ移り、音色の変化を観察する
- 一般に F1 は低く、F2 は高くなる。**音程の主な手がかり F0 とは別**。実際の周波数の変化は後の回で分析して確かめる

演習3：マクガーク効果の動画鑑賞

- 映像あり／音声のみで比べ、聞こえた音をチャットに記録する。効果の出方には個人差がある

次回（第3回）の予告

形式：動画回（非同期・自分のペースで視聴）
テーマ：音をコンピュータで扱うための数値化
内容：サンプリング・量子化・ナイキスト定理・WAVファイル・Pythonで音声を読む

事前学習（任意推奨）

1. 手元の音声ファイルの情報表示で、サンプリングレートやチャンネル数を探す。
   表示項目はアプリや形式によって異なり、見つからなかった項目はそのままでよい。
2. 「なぜCDは44.1kHzなのか」を検索してみる

第2回のまとめ

1. **エネルギー源・音源・フィルタ**を区別する。呼気がエネルギーを供給し、声帯振動などが音源となり、声道が周波数成分を変える
2. 子音は調音部位・調音様式・有声／無声などで分類する。**F0 と F1・F2**を区別する
3. 耳の蝸牛は周波数に応じて異なる場所で応答し、振動を神経系の信号へ変換する
4. 音声知覚には、聴覚だけでなく**注意・記憶・文脈・視覚**などが関わる
5. ASRは音声を文字列へ変換する技術。雑音・文脈・視覚情報の扱いはモデルにより異なり、人間と比較するには課題と評価条件をそろえる

小課題：翌週末23:59 までに Google Drive または メールで提出（400〜600字程度）