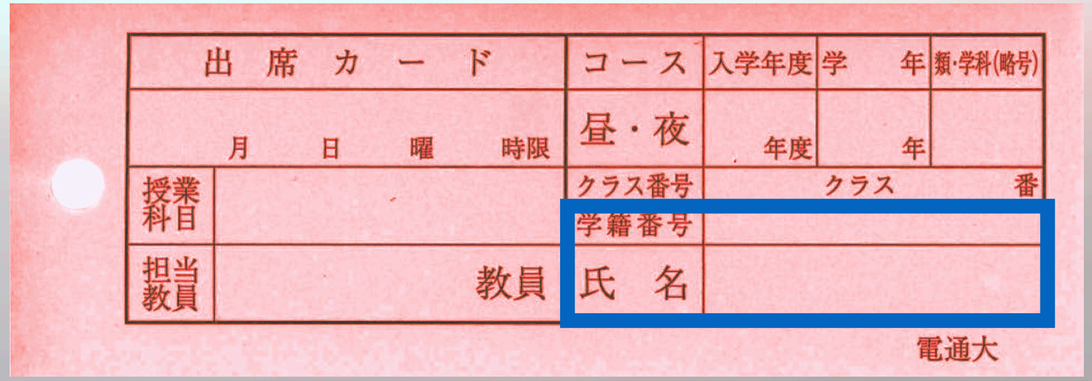
メディア情報学実験
メディア情報検索・認識実験（音声）

座席は自由

ubuntu で実習を行います

できるだけ向かって右側に座ってください
左端付近は隣の実験の説明の音がよく聞こえます



出席カード		コース	入学年度	学　年	類・学科(略号)
月　日　曜　時限		昼・夜	年度	年	
授業科目		クラス番号	クラス　番		
		学籍番号			
担当教員	教員	氏　名			

電通大

ここだけ記入すればよい
数字は丁寧に書いてください

実験資料　IEDホームページの「教材」からリンクあり

www.takagi.inf.uec.ac.jp/swr/

電気通信大学 I類／II類セキュリティ情報学プログラム 教育用計

概要　活動　**Wiki**　運用スケジュール　はじめて利用する方へ　時間割表　FAQ（よくある質問）

はじめに

このWebサイトは、I類／II類セキュリティ情報学プログラム 教育用計算機演習室（IED演習室）のサポートページです。
計算機室をはじめて利用される方は、 **はじめて利用する方へ** を参照してください。
頻繁に寄せられる質問は **FAQ（よくある質問）** にまとめてあります。

お知らせ

2022/7/28

- Windows へのログイン時に「Windows のライセンスの有効期限がもうすぐ切れます」と言う警告が表示される問題が
  の警告は無視して、メッセージウィンドウを閉じて下さい。

2022/4/7

- システム更新に伴い、端末の遠隔利用方法が大幅に変更されていますのでご注意下さい。24時間運転する端末はなく、
  グインする方式に変わっています。

2022年度より、計算機の端末がリニューアルされました。

- 従来システムからの主な変更点
- 演習室端末(Linux)の遠隔利用
- リモートデスクトップによる演習室端末（Linux）の利用
- 運用スケジュール
- IED 演習室は飲食物の持ち込みを禁止しています。飲み物を持ち込みカーペットを汚す事例が多数出ています。
  持ち込みを確認した場合は掲示板で警告しているように2週間アカウントを停止します。
- 管理の都合上、 **ディスプレイの電源はオフにしない** でください。

講義・演習

- 教材
- 電気通信大学 WebClass
- 時間割表

施設案内

- システム概要

教材

I類

- 情報領域演習第２ C演習 クラス1
- プログラミング言語実験
- プログラミング演習
- 数値解析
- メディア情報学実験 メディア生成・制作
- メディア情報学実験 メディア基礎
- メディア情報学実験 メディア情報検索・認識（音声）

II類

- プログラミング言語実験（松本、市野、山本）

旧情報通信工学科

- アルゴリズム基礎論演習（昼）
- アルゴリズム基礎論演習（夜）
- プログラミング演習
- 計算機言語演習
- 情報通信工学実験A・B
- 基礎プログラミング演習

上記にリンクを追加したい場合は、Web 教材と実験用データ をご参照ください。

▲サポートページトップに戻る

メディア情報学実験
メディア情報検索・認識実験（音声）

第 1 章 はじめに　1.7 計算機

1.7　計算機実習の準備

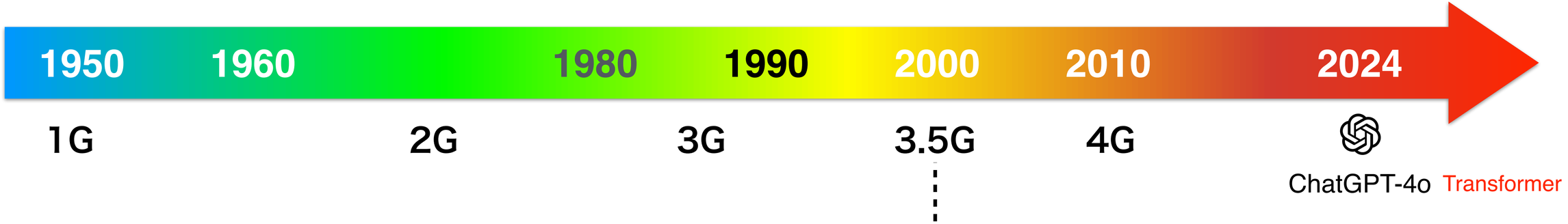
1.7.1　ファイルのコピー

自分のホームディレクトリの直下に asr というサブディレクトリ
験に必要なファイルを用意します.

```
[~]% cd
[~]% cp -pR ~takagi/asr .
```

まずは、とりあえずこれを実施しておく

音声認識の歴史



基盤技術の高度化：誤り率最小化に基づく識別的モデル
モデル適応，話し言葉音声DB

1G 音声信号の物理的特徴に関する知見
アナログフィルタバンクと論理回路を用いた経験則的な処理

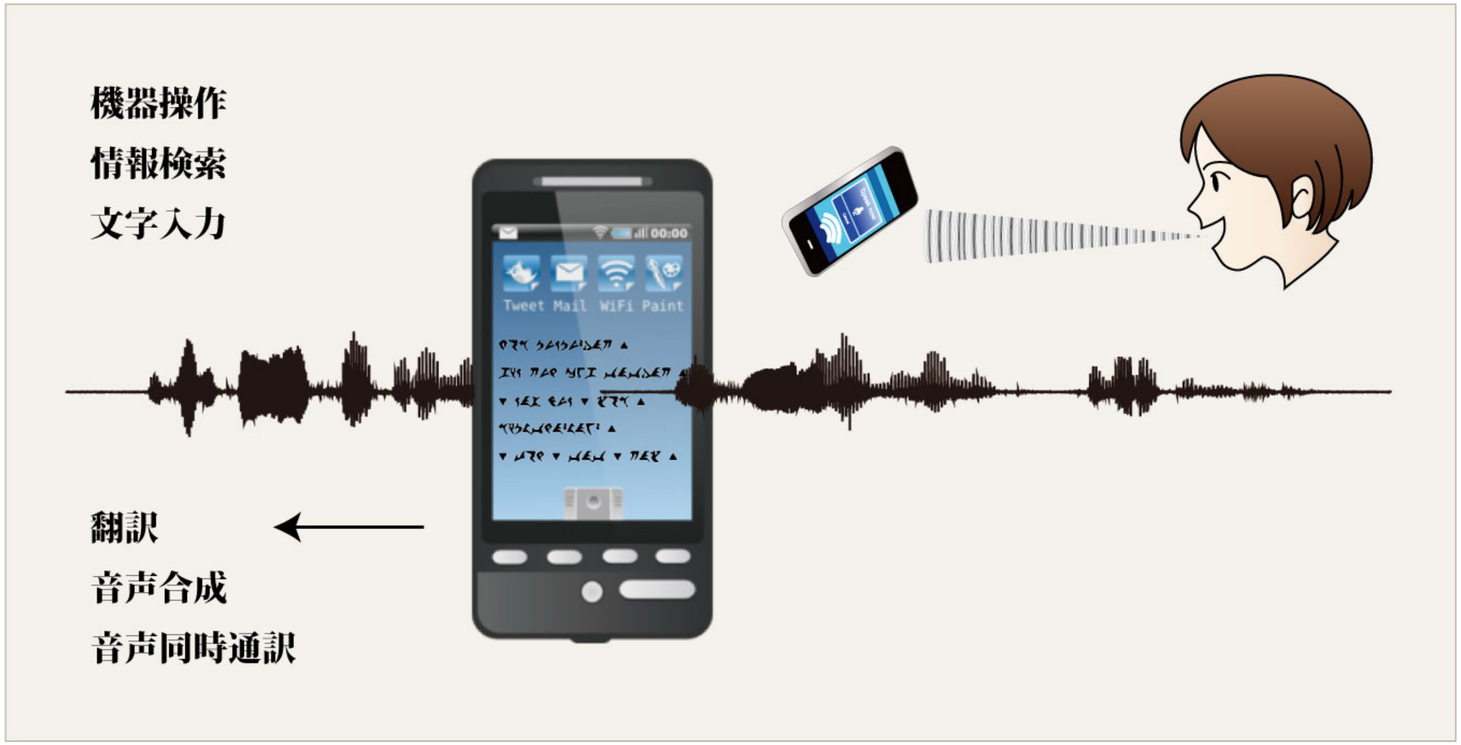
2G 音声認識に有効な音響特徴量と動的パターンマッチング手法に関する研究が進展
テンプレートとDPマッチングに代表される時間軸整合マッチング

3G GMM-HMM：DPマッチングを拡張した形で隠れマルコフモデル（Hidden Markov Model: HMM）
HMMの各状態の音響特徴量のパターンをモデル化する混合正規分布（Gaussian Mixture Model: GMM）
N-gram言語モデルの導入によりその後の実用的音声認識システムの基盤技術になる

4G DNN-HMM：深層ニューラルネットワーク（Deep Neural Network: DNN）による大幅な性能向上
音響モデル：DNN-HMM
言語モデル：Recurrent Neural Network（RNN）とN-gramの併用
RNNを発展させた長・短期記憶（Long Short-Term Memory: LSTM）のみでEnd-to-End処理
Transformerを中核としたマルチメディア情報処理の革新

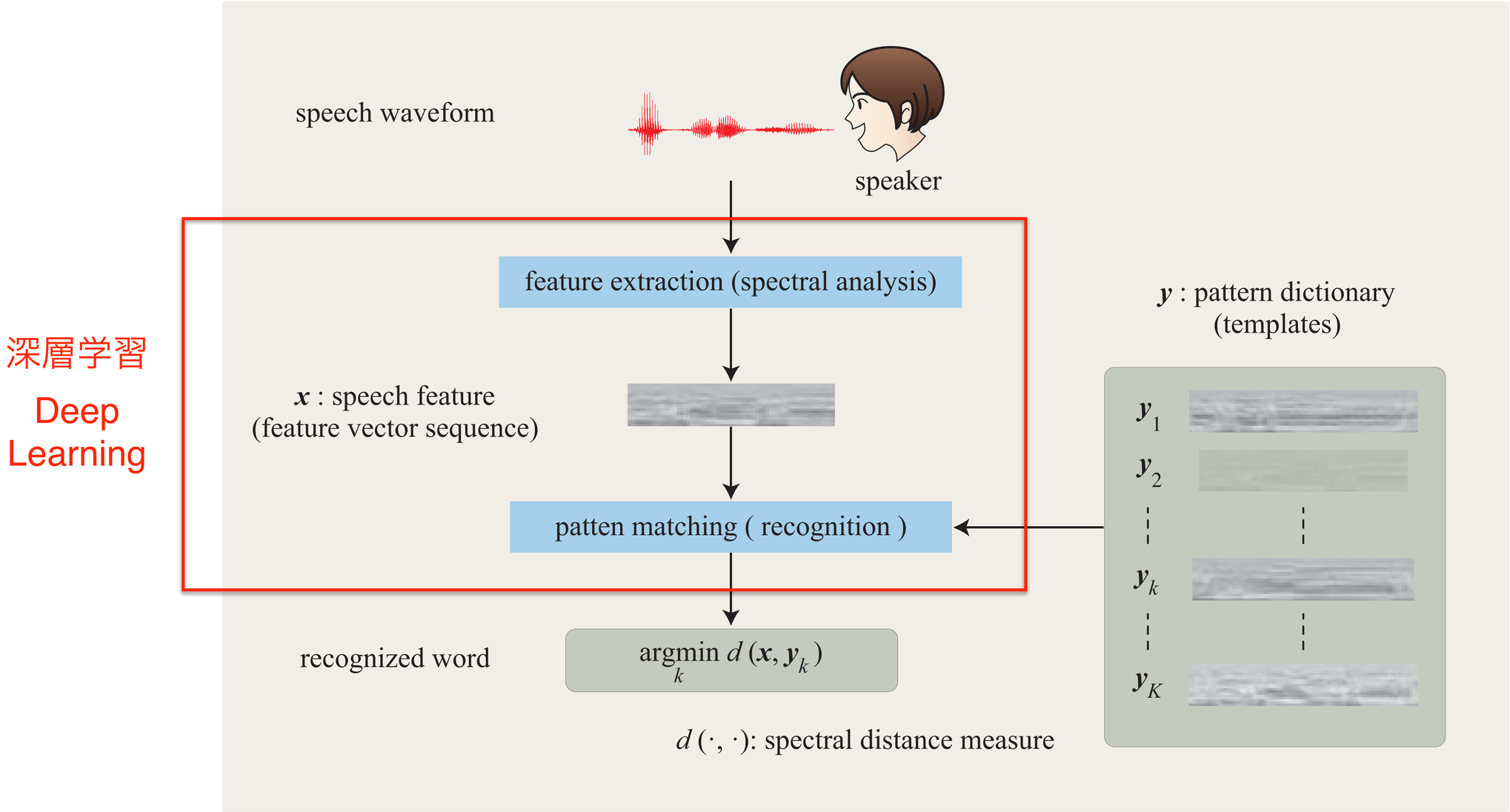
はじめに

音声のパターン認識



携帯端末やスマートスピーカーの音声入力機能を用いて、
テキストメッセージを入力したり，道順を調べたり，
ウェブ情報の検索をすることが実用化されている。
音声自動翻訳も…

音声のパターン認識



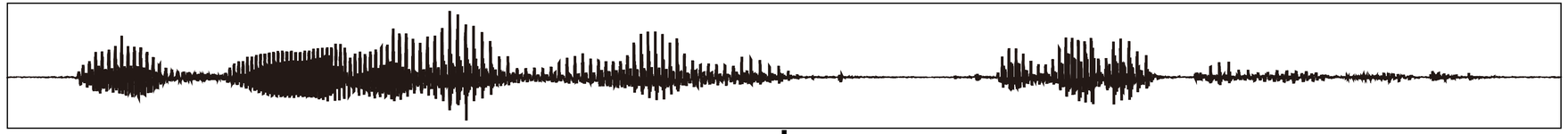
コンピュータに登録した認識対象の単語の音声パターンの中から
入力音声のパターンに最も似ているものを選んで認識結果として出力。

音声認識はパターン認識で最も難しい範疇

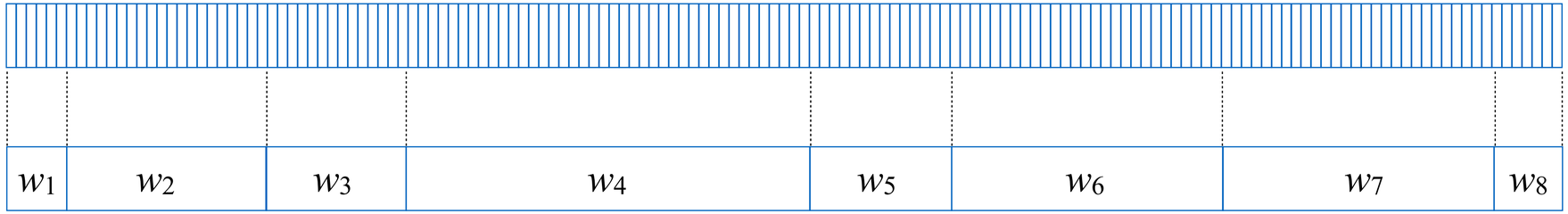
1出力1入力

→ 手描き文字認識 → 2

複数入力複数出力
系列長対応関係無



音声認識：単語系列推定

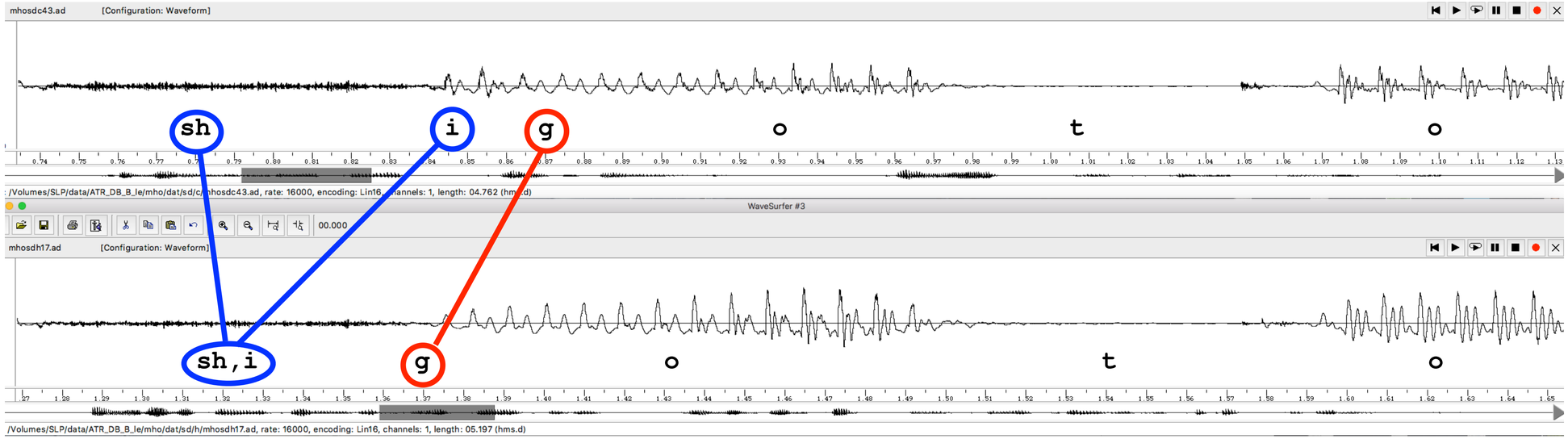


・入力は複数（一定の時間間隔で区切られた特徴量の系列），出力は単語系列。

・入力の系列長と出力の系列長の間に明確な対応関係無し。

・早口で話すのとゆっくり話すのとで，単位時間の単語数が違う。

同じ単語＝同じ波形？



ATR研究用日本語音声データベース

話者= mau

単語＝仕事　/shigoto/

上段：主人の仕事は建築業で作業服を着て出勤します（C43）

下段：限られた予算で仕事をする美術スタッフの気持ちをおしはかって我慢する（H17）

同一話者が同じ単語を発声しても、同じ波形にはならない。

個々の音（音素）の継続長、タイミング、音色（スペクトル）が違う。

実習日程

第１週

- 第１章　はじめに
- 第２章　音声
- 第３章　スペクトル分析
  - フィルタバンク分析　C プログラム
  - MFCC 分析　C プログラム
音声の録音と編集

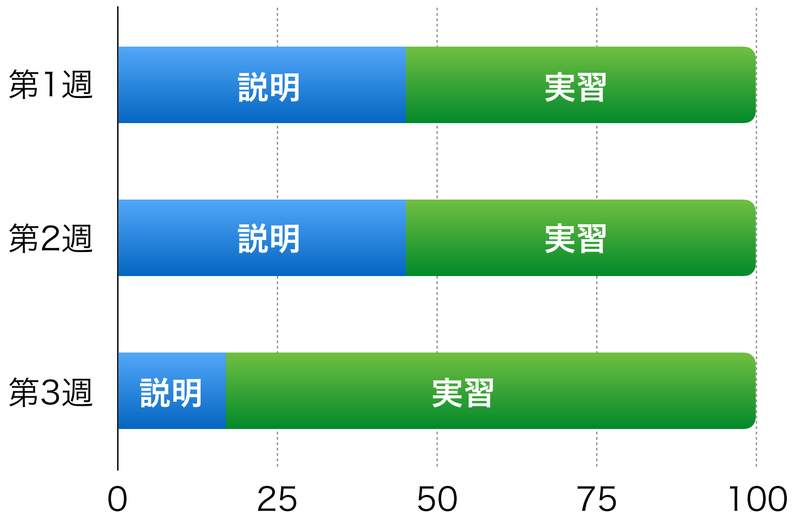
第２週

- 第４章　統計的音声認識
- 第５章　隠れマルコフモデル
  - Forward アルゴリズム　C プログラム
  - Backward アルゴリズム　C プログラム
  - Baum-Welch アルゴリズム　C プログラム
学習と認識の
シミュレーション
  - 多次元ガウス確率密度関数　C プログラム

第３週

- 第６章　単語認識
  - 音声データ作成（収録・編集）　音声入力
ラベル付け
  - 単語 HMM の学習・性能評価　データの一括処理
シェルスクリプト
  - On-The-Fly 単語音声認識　単語認識の評価

時間配分見込



音声

二重文節

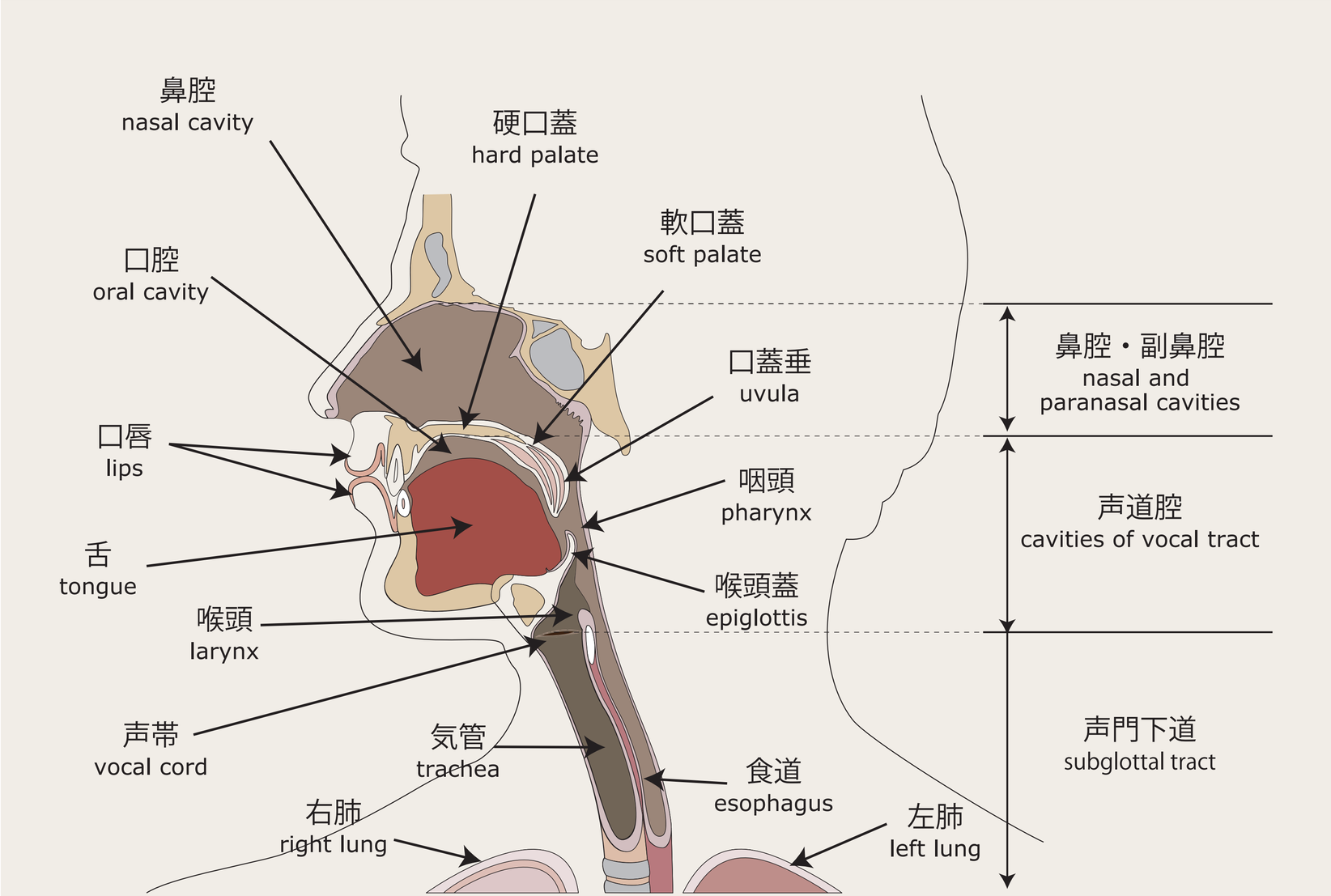
文	太 郎 は 学 校 へ 行 っ た				
形態素（語）	太郎	は	学校	へ	行った
音素	t a r o o w a g a Q k o o e				i Q t a

・音素が組み合わされて形態素（語）を構成し，形態素（語）が組み合わされて文が構成されるという二重構造。

・比較的少数の音素を任意の個数並べることにより数千〜十数万の単語を構成することができる。

・単語を文法に従って任意の個数並べることによって，事実上無限種類の文の生成が可能。

発音器官概略図（頭部正中矢状断面）

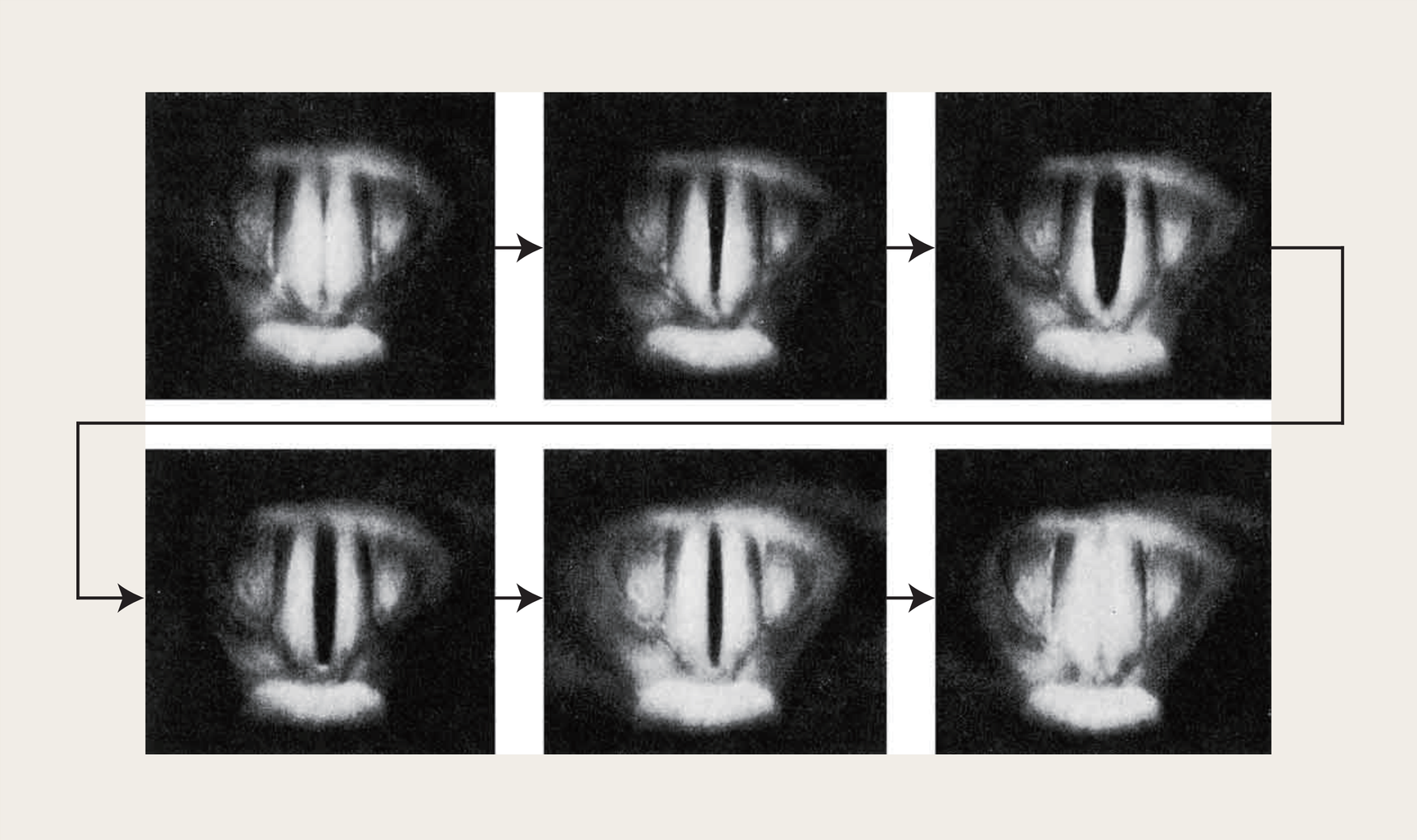


音声器官は肺，喉頭，咽頭，口，鼻から成る。

声道は咽頭，口からなり，その形は母音の種類によって異なる。

声道の形は口唇，顎，舌，喉頭の位置や形によって変化

声帯 vocal code

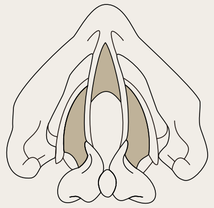


声帯の振動の様子（1 周期，約8 ms)

適切な張力が掛かった声帯の間を呼気が通ると，声帯は細かく振動し，
断続的な空気流を声道に送り込む。

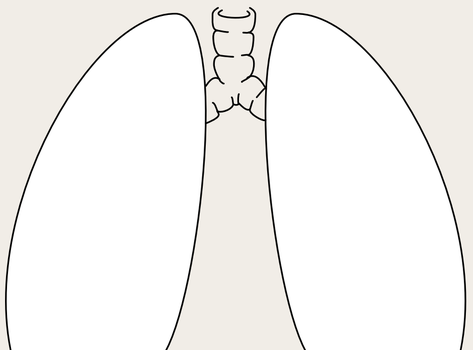
音源
Source

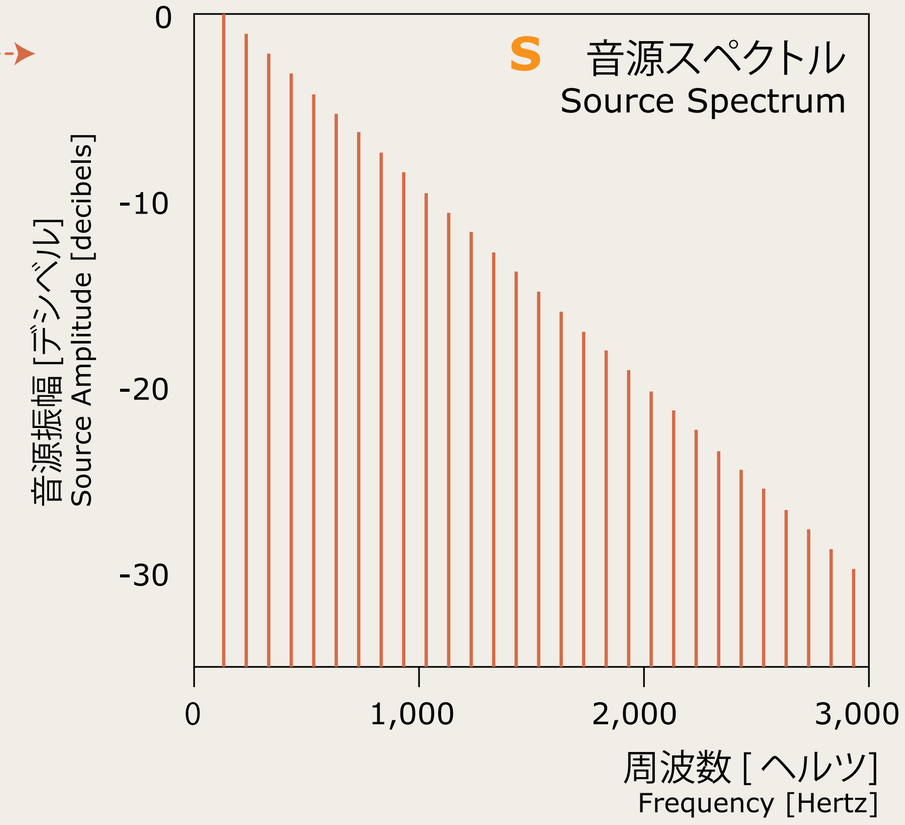
声帯振動
Vibrating Vocal Folds



空気流
Airstream

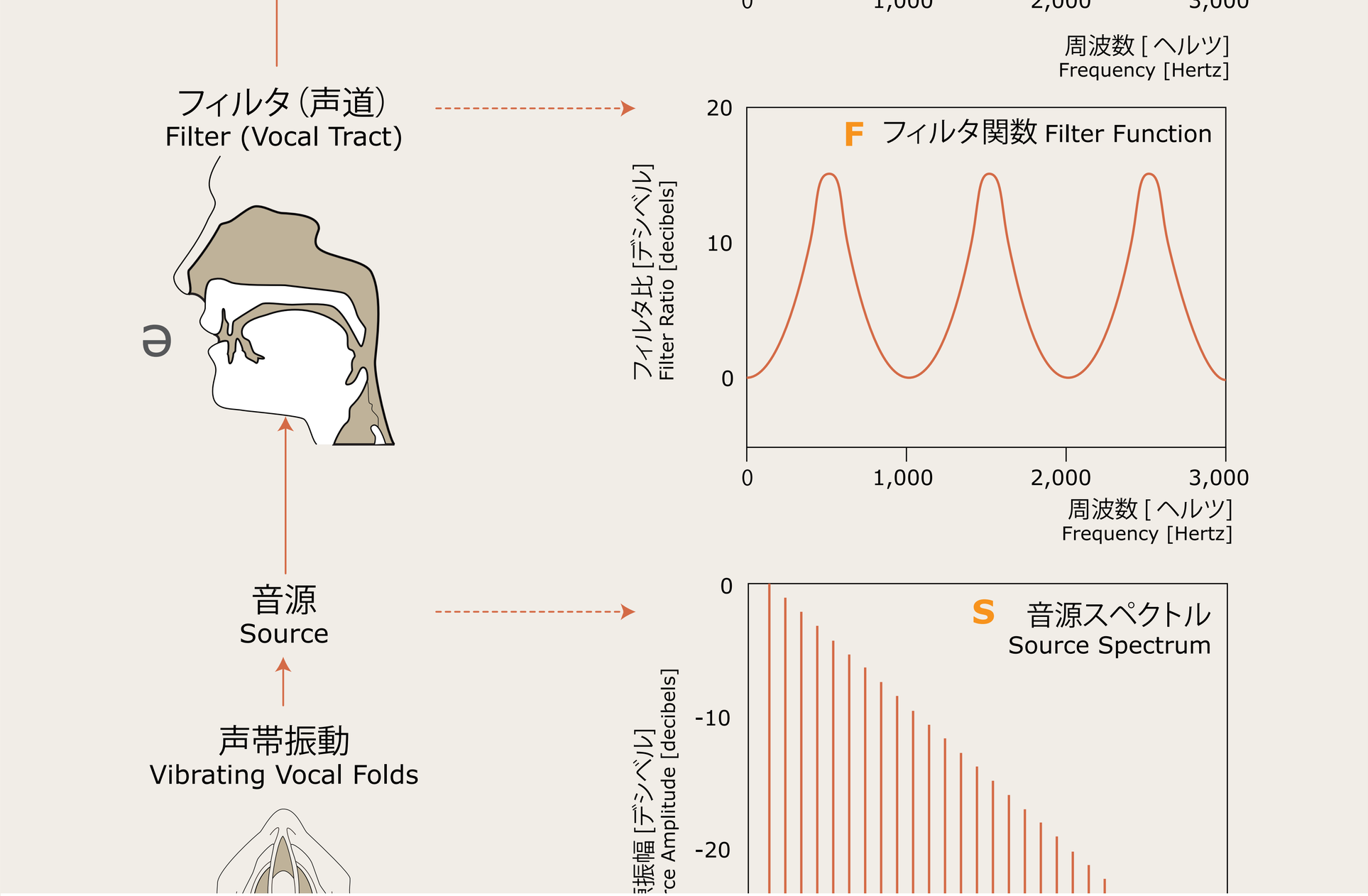
肺
Lungs





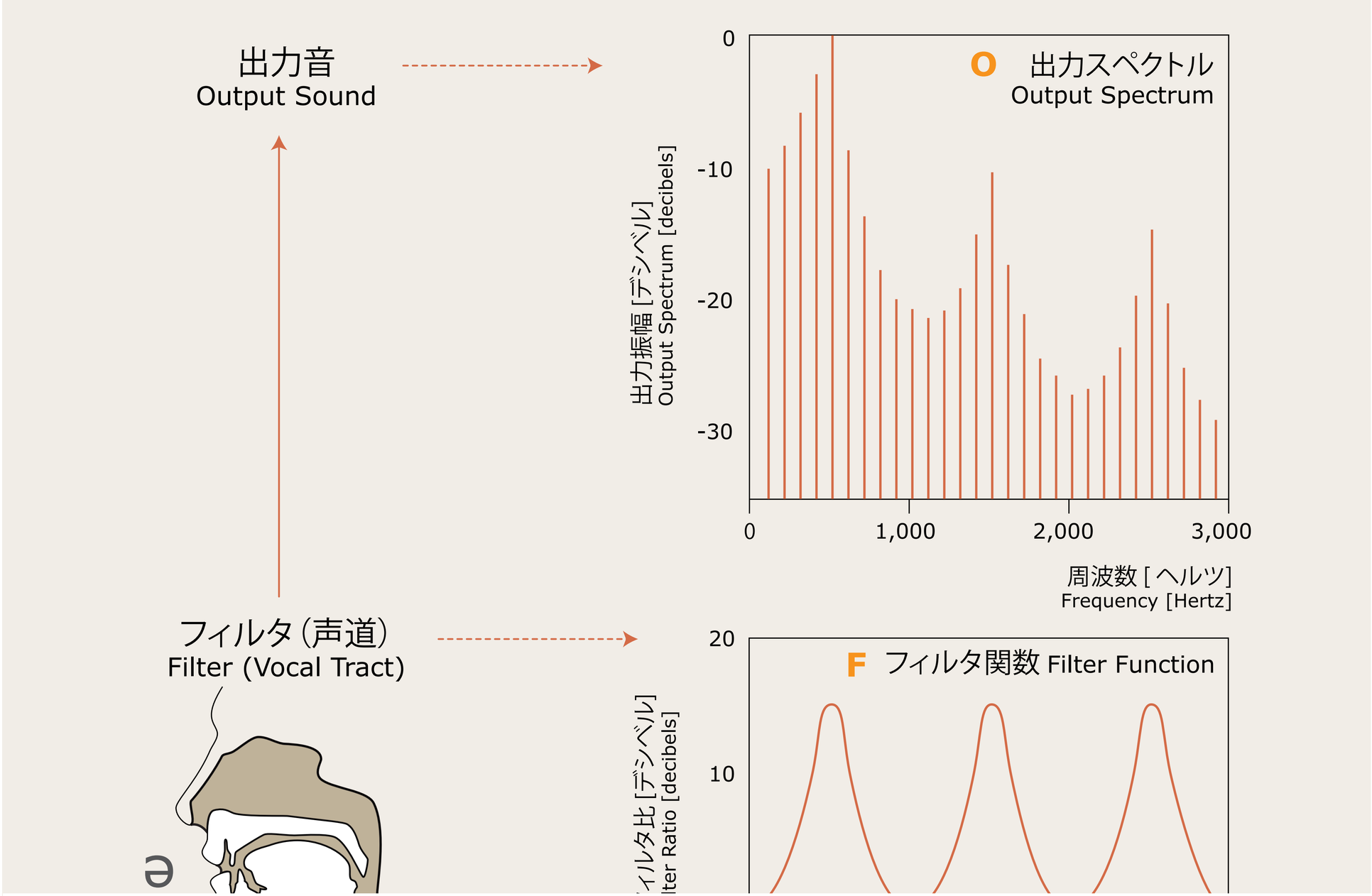
・音源は肺からの空気流によって振動する声帯によって発生する声帯音源。

・音源の音響スペクトル**S**は周波数が高くなるに従って振幅が減少（-12dB/oct）。



・音源・フィルタ理論のフィルタに相当するのは声道。

・音源の周波数特性に変化を与え、その結果フィルタ関数 **F** が得られる。

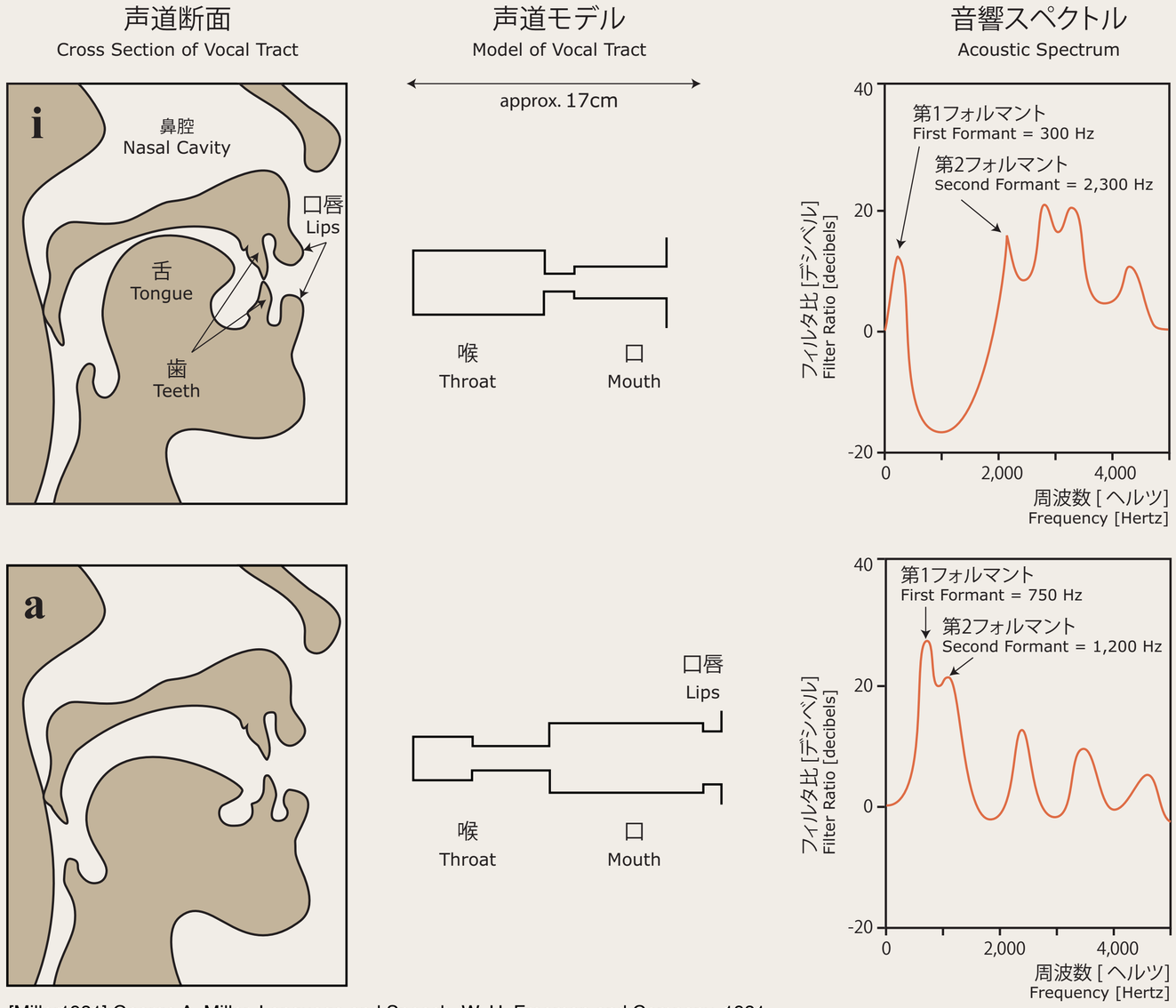


音源にフィルタが適用されると，母音の出力スペクトル **O** に共振ピークが生じる。

O = S × F (周波数領域)。

声道モデル

- MRI画像等から得られた発声器官の形状から声道断面積を推定し、それを用いて声道のフィルタ特性を計算。
  - 左：母音の声道断面の概形。
  - 中：少数の円筒を接続した声道モデル。
  - 右：声道モデルから計算した音響スペクトル。
- 聴覚実験により、人間は音響スペクトルの概形（特に共振周波数）によって、母音の種類を識別していることが分かっている。
- 声道の共振周波数は母音の認識にとても重要なので、特別に**フォルマント**と呼ばれている。



[Miller1981] George A. Miller, Language and Speech, W. H. Freeman and Company, 1981.

声道模型教材

VTM-10 NTTアドバンステクノロジ

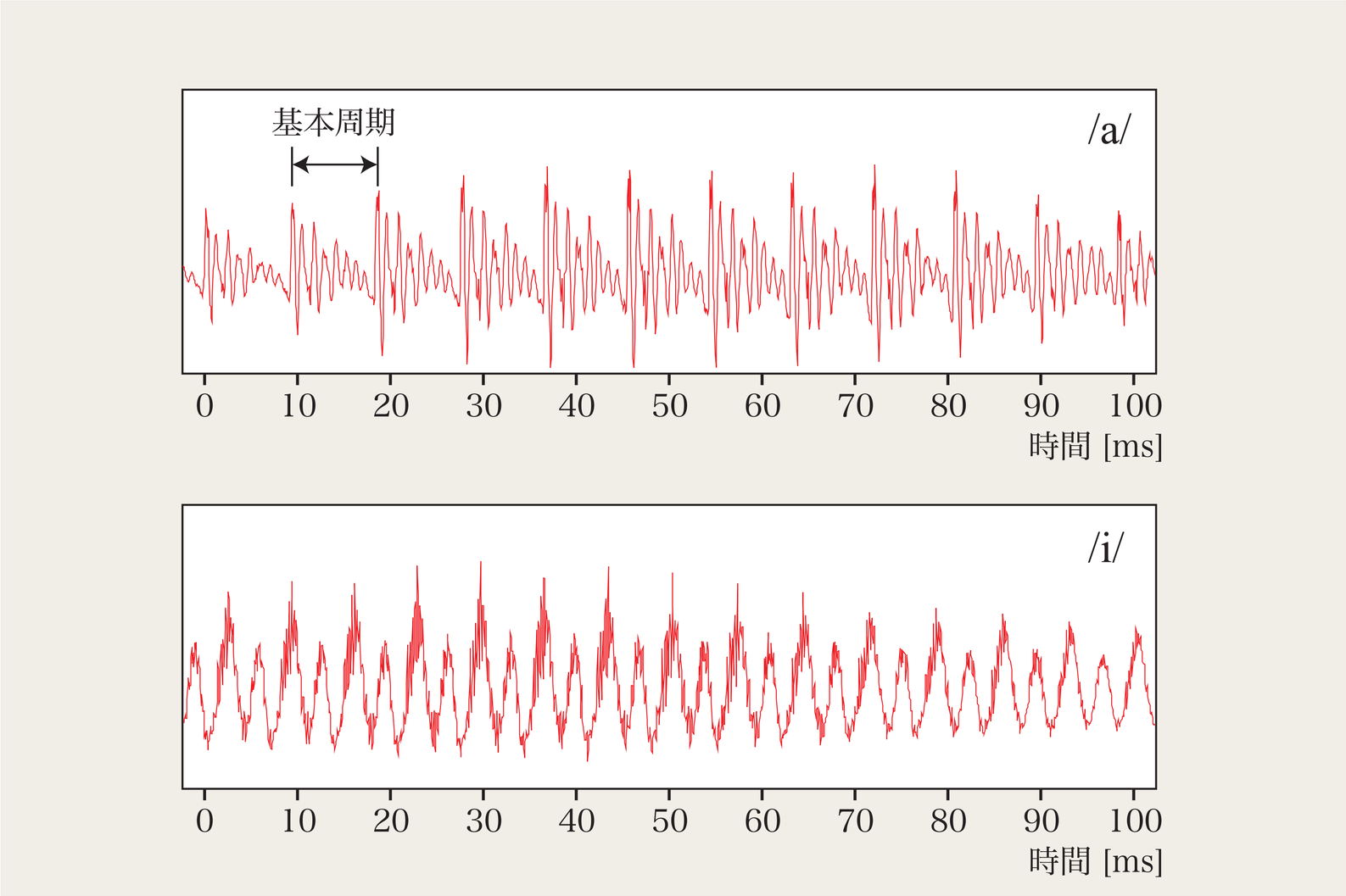
声帯音源

SECOM
電動式人工喉頭
マイボイス

声道フィルタ

パイプ型模型
直径70mmアクリル樹脂

日本語母音の音声波形（男性）

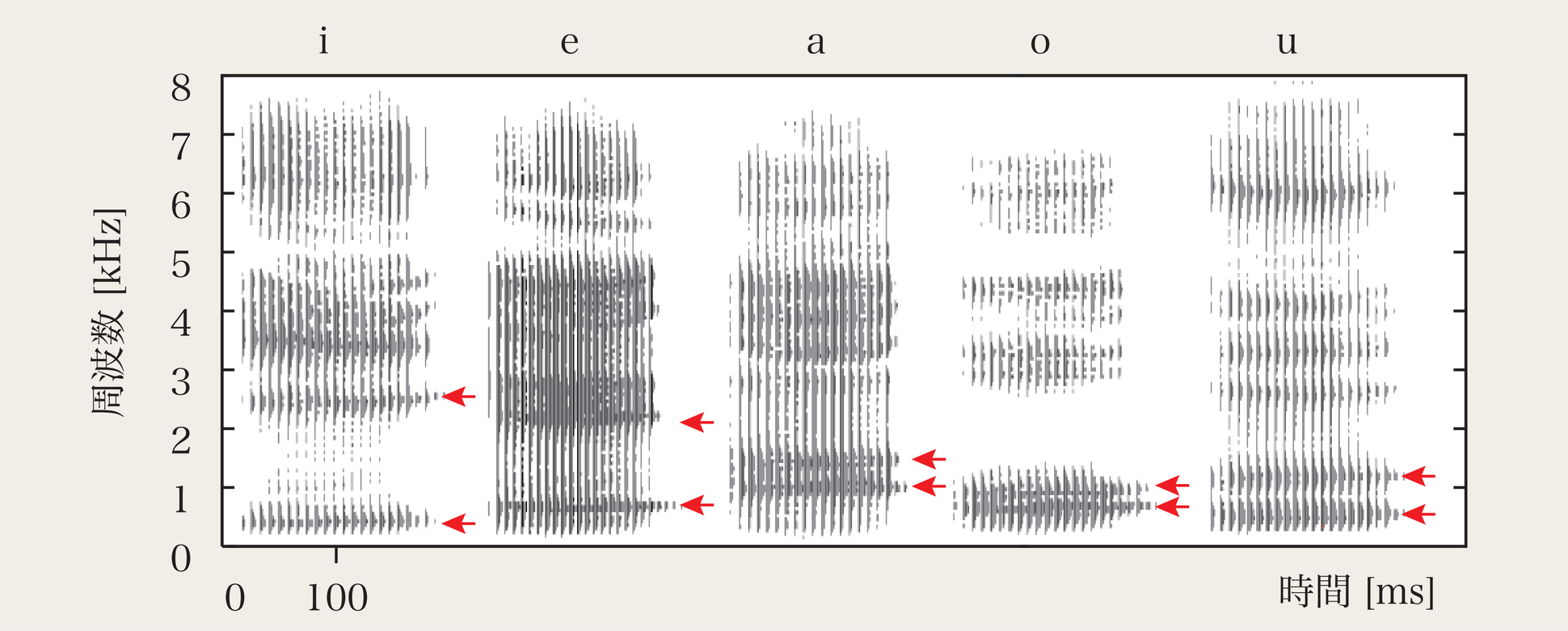


・音声波形は母音によって形が大きく異なる。

・同じような波形の繰り返し。この/a/ の場合は約9ms。

・この周期を音声の基本周期，基本周期の逆数を基本周波数という。

日本語５母音（男声）のスペクトログラム表示



- 音声スペクトルには音素によって特徴的な山や谷。山の部分を**フォルマント**という。その代表周波数（山のピークの周波数）をフォルマント周波数という。
- 周波数の低い方から第1フォルマント、第2フォルマント…と呼び，F1，F2と表記。
- この図の濃い横筋がフォルマントに対応。
- F1とF2 の位置（矢印）は母音によって明確に異なる。

母音はF1–F2平面で分離可能

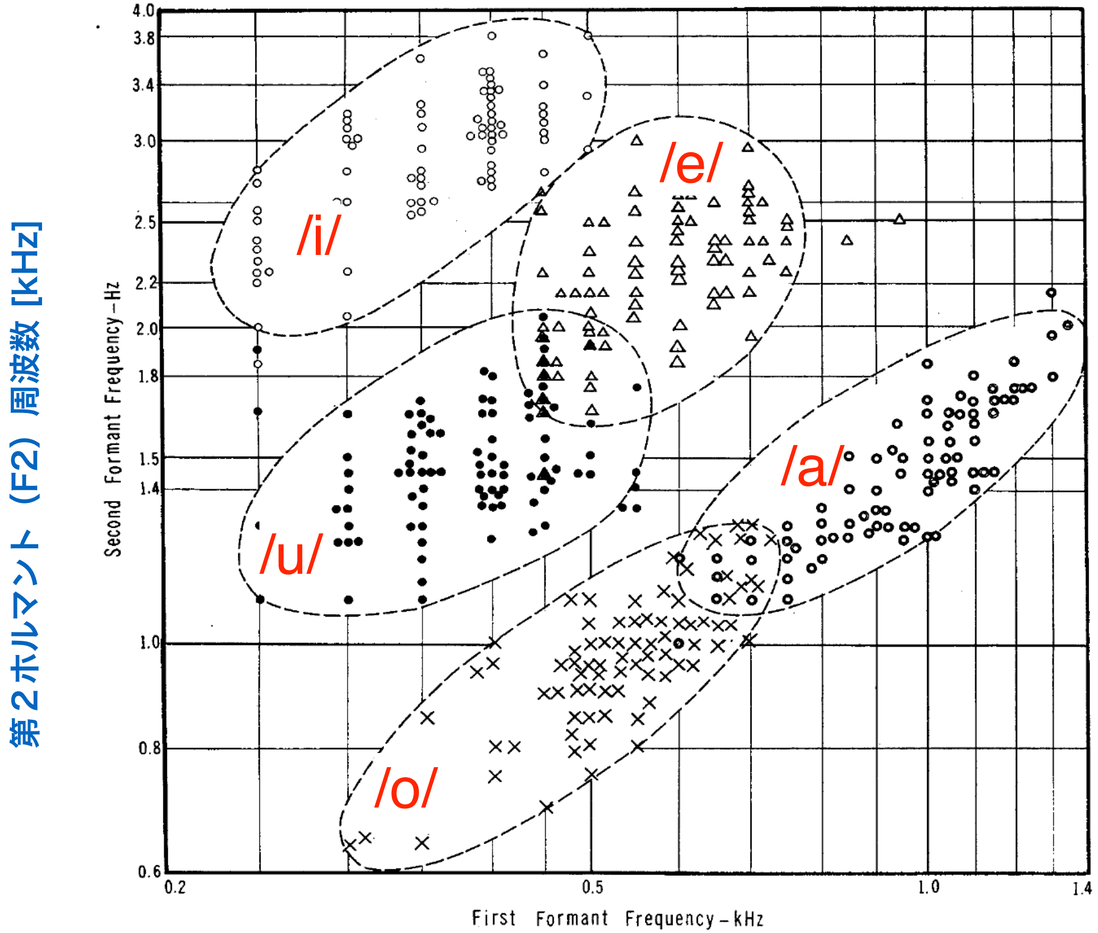


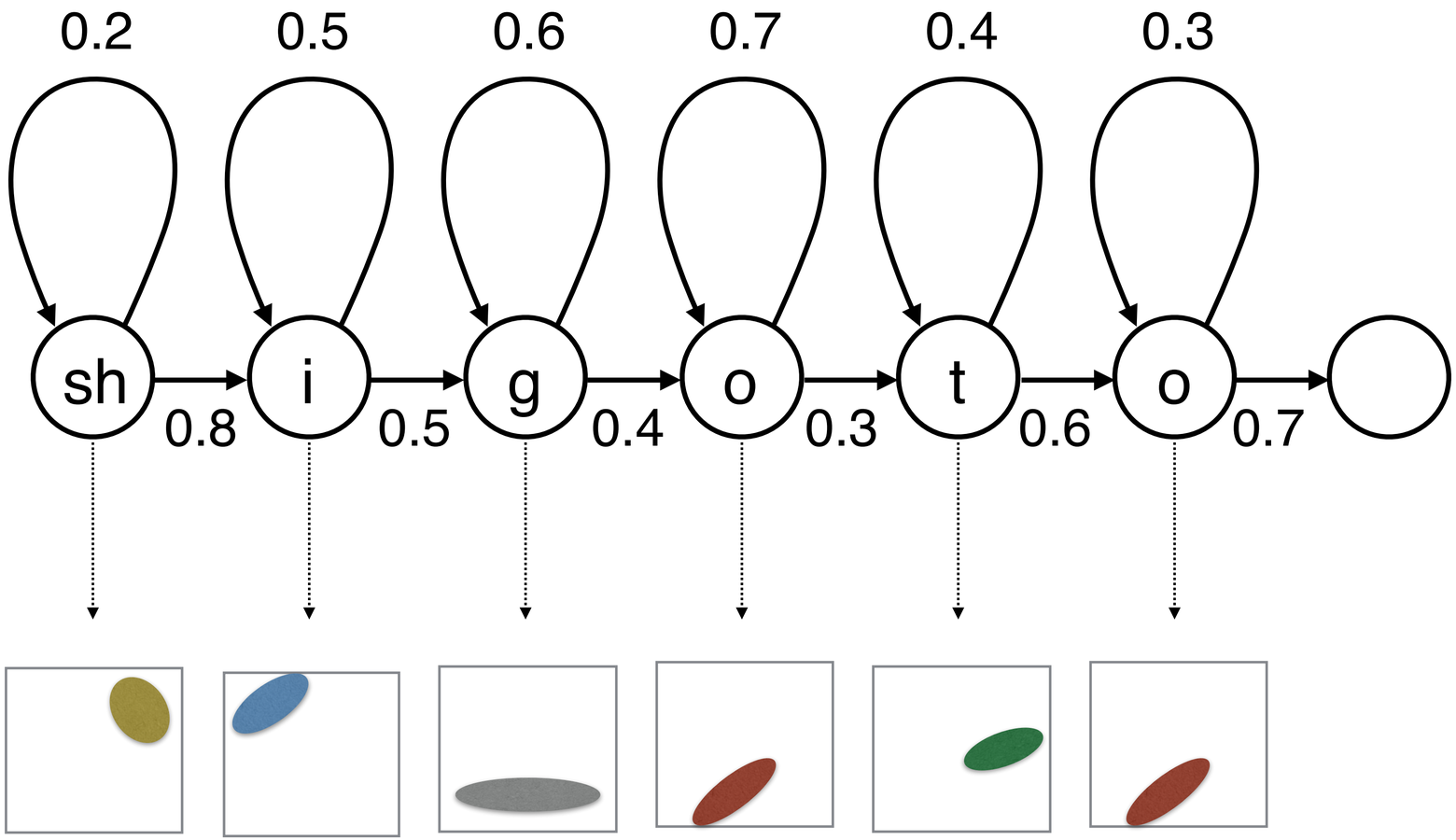
Fig. 8 Second formant frequency (F_2) versus first formant frequency (F_1) diagram of five Japanese vowels spoken by children, youth and adults.

第１ホルマント（F1）周波数 [kHz]

日本語５母音の第１ホルマント周波数に対する第２ホルマント周波数（子供、若年者、成人）

出典：粕谷英樹，鈴木久喜，城戸健一，”年令，性別による日本語5母音のピッチ周波数とホルマント周波数の変化”，日本音響学会誌，Vol.24，No.6，pp.355–364，1968年，11月.

Hidden Markov Model 隠れマルコフモデル



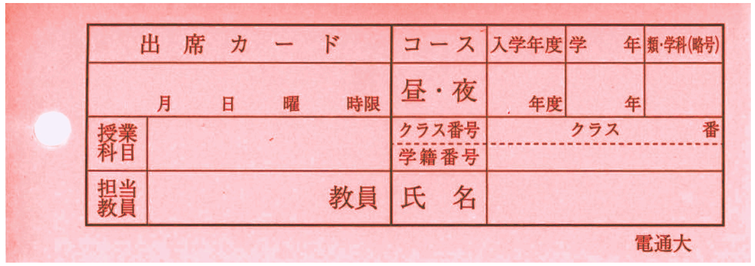
・単語中の音（音素など）の時間変化を状態遷移でモデル化

・音素のスペクトルの分布を状態出力分布でモデル化

出席票回収、ヘッドセットと実験手順書配布

- 「出席カード」と交換
- 教室の前方に取りに来てください。
- パッケージ、結束バンド等は各自全て持ち帰ること。
  IED入口のゴミ箱には絶対捨てないでください。
- ヘッドセットの不良品は交換します。本日の時間中に動作確認してください。

回収



配布

ヘッドセットの種類…アナログ接続3.5mm4極プラグ または USB

ELECOM社製

実績有

HS-HP22TBK
（3.5mm４極プラグ接続）
無指向性マイク

今年度試用初導入（USB接続）：希望者のみに配布します

× 5

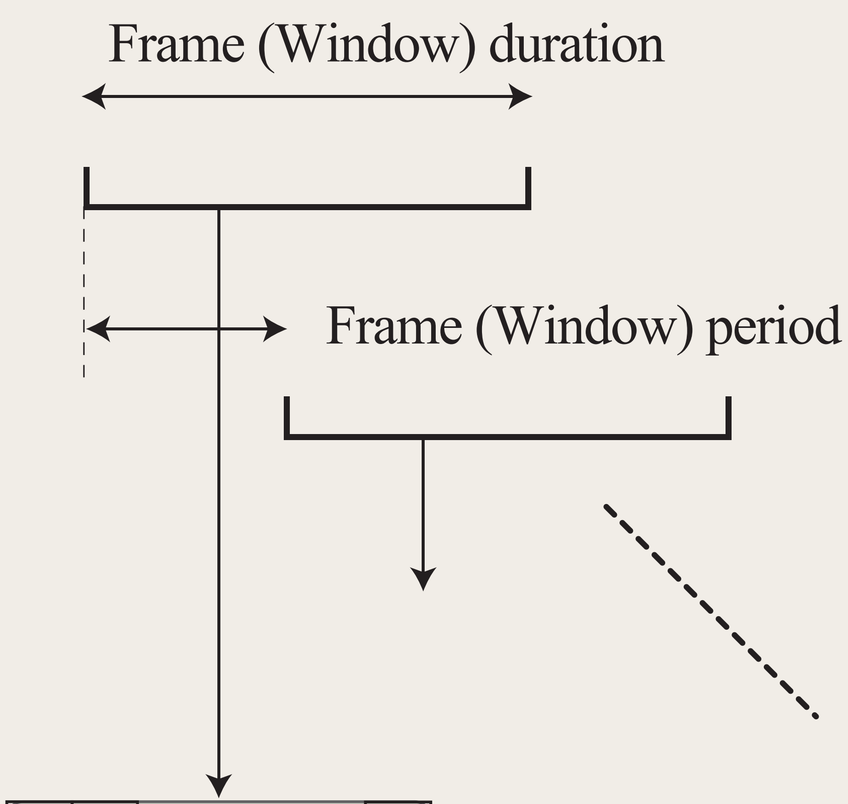
HS-HP07SUBK（USB接続）
ヘッドフォン部大型
指向性マイク（ノイズ低減）
指向性があるので入力音量確保に注意

× 5

HS-HP14SUBK（USB接続）
無指向性マイク

スペクトル分析

speech waveform $\boldsymbol{s}$



短時間フレーム分析

- 音声波形から連続する短い時間区間を切り出して分析。
- 切り出す単位をフレームという。
- フレームを一定幅で時間軸方向にずらし、切り出した音声波形データからスペクトル特徴量（特徴ベクトル）を求める
- この処理をフレームが音声波形の終端に達するまで繰り返す。

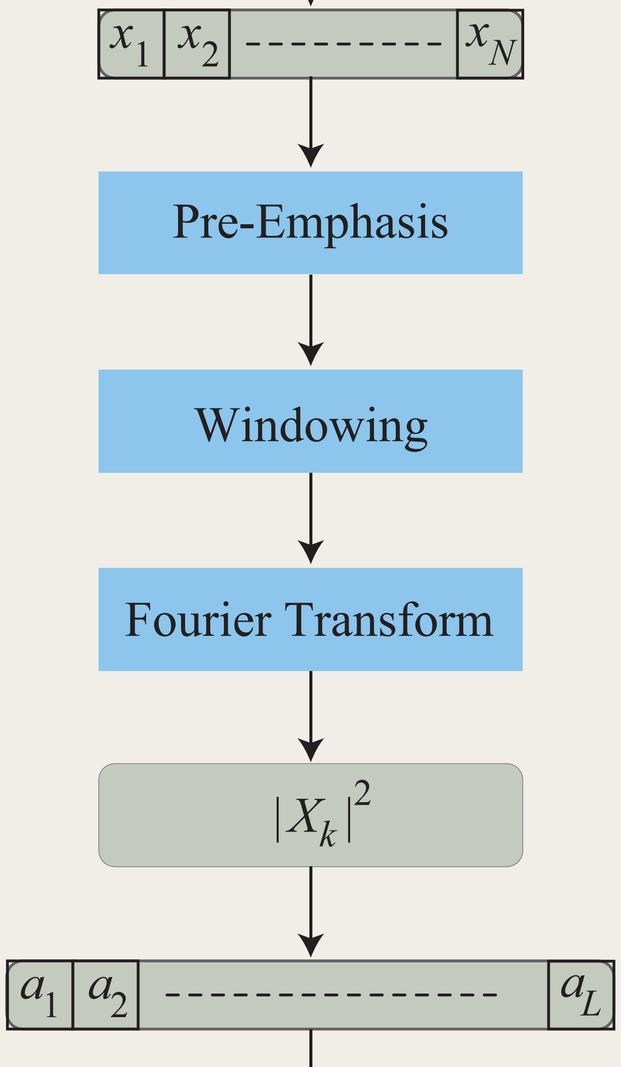
Speech Samples: $\boldsymbol{X} = x_1, \ldots, x_N$

$$H(z) = 1 - az^{-1}, \quad x_n \leftarrow x_n - ax_{n-1}$$

高域強調　音韻性情報の強調　声帯音源のスペクトル傾きを低減

窓掛け　フレーム両端の不連続の影響の低減

フーリエ変換　スペクトル成分に分解



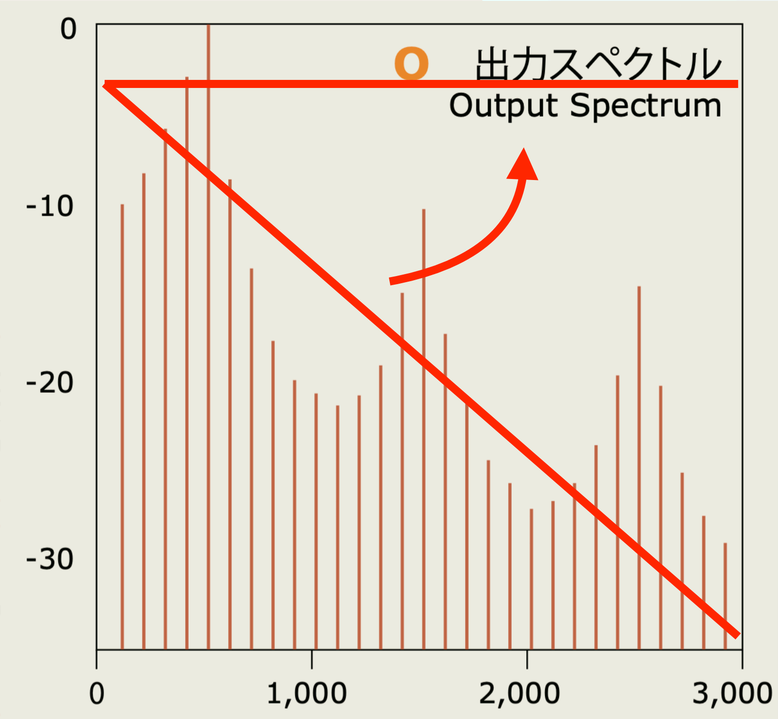
Speech Samples: $\boldsymbol{X} = x_1, \ldots, x_N$

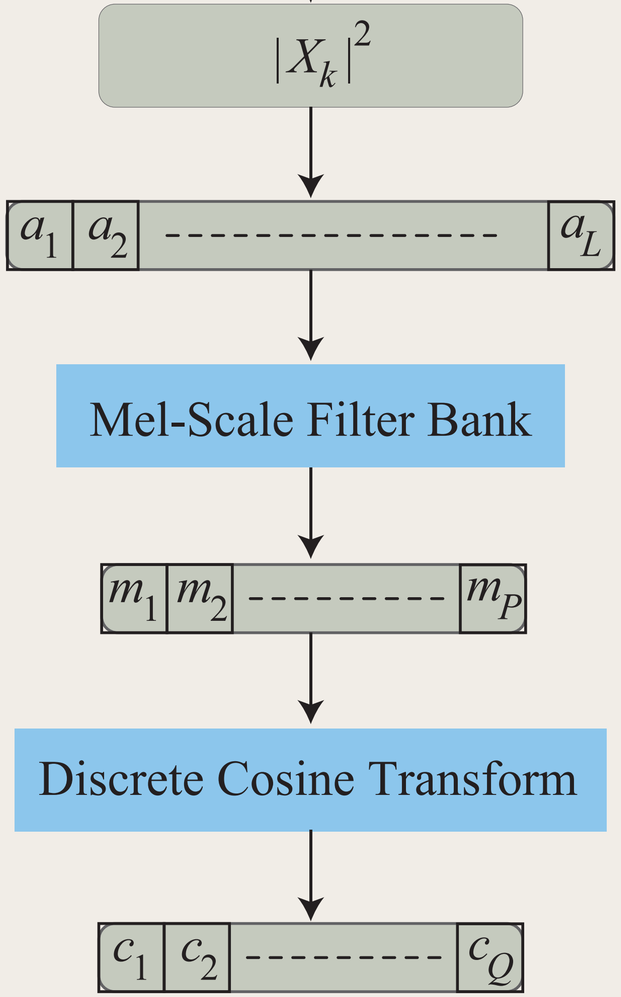
$$H(z) = 1 - az^{-1}, \quad x_n \leftarrow x_n - ax_{n-1}$$

$$x_n \leftarrow w_n x_n, \quad w_n = 0.54 - 0.46 \cos \frac{2}{N}$$

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j\frac{2\pi nk}{N}}, \quad k = 0, 1, \ldots, L$$

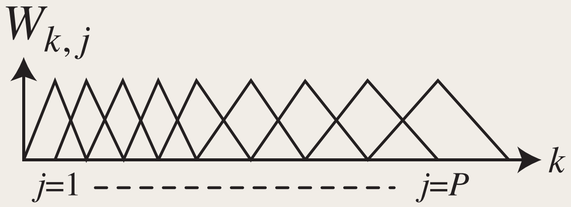
DFT Power

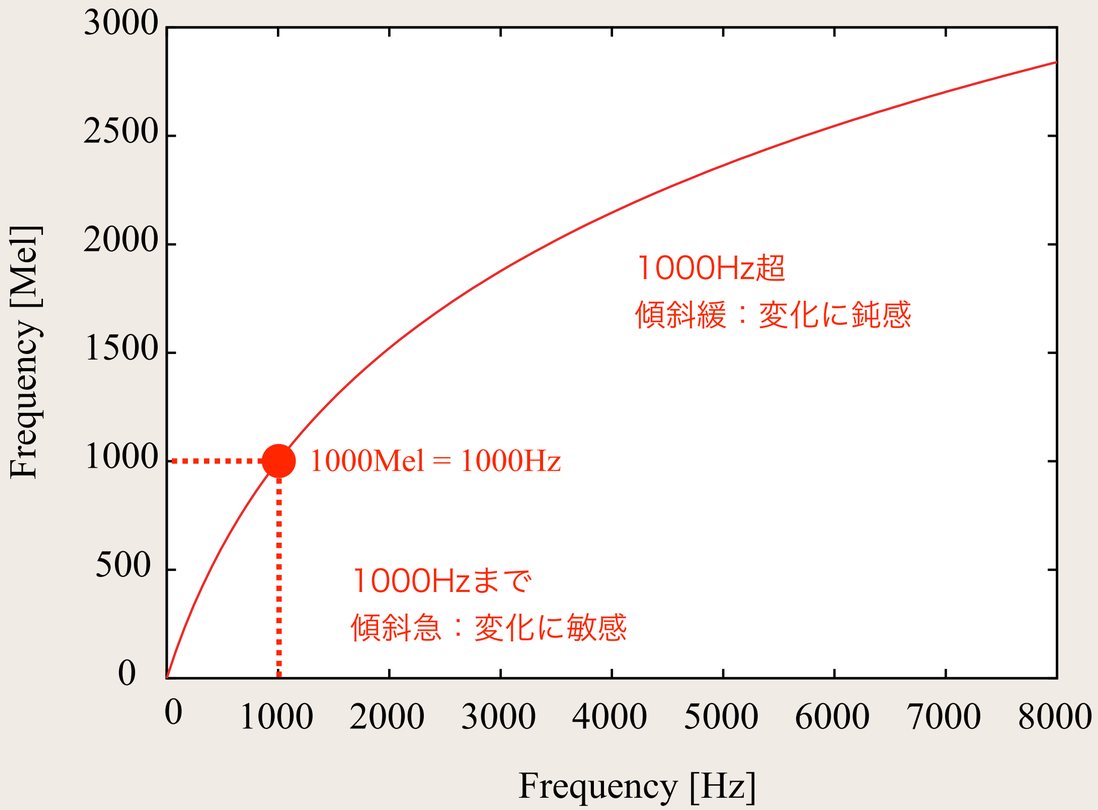


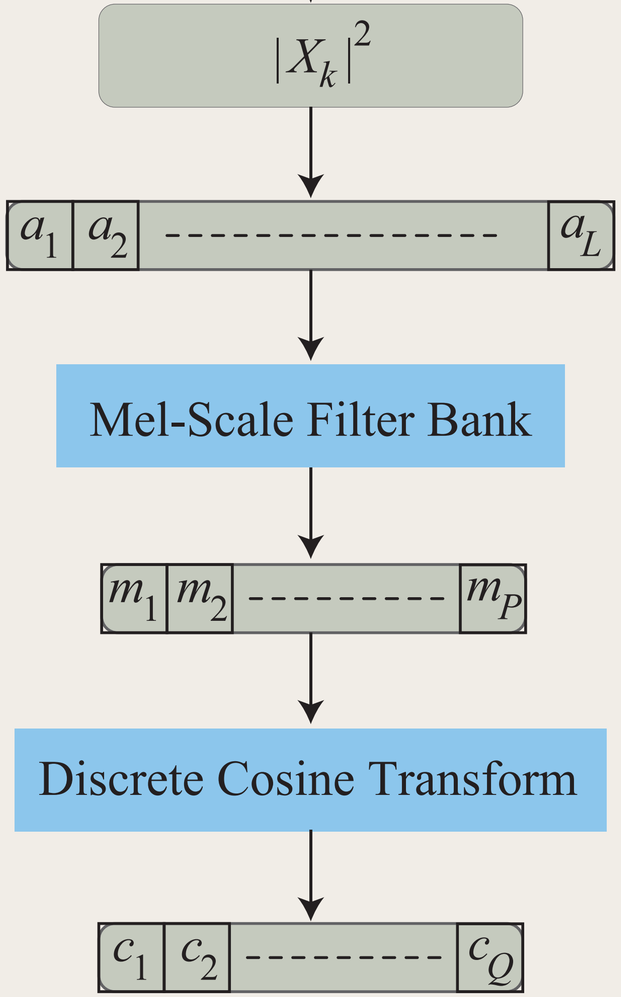


メルフィルタバンク分析

- 聴覚の周波数分解能は、低い周波数では高く，高い周波数では低い。
- 周波数が高くなるほど幅が広くなる三角窓が並んだフィルタバンクで処理。

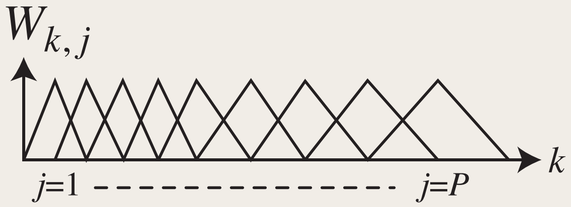






メルフィルタバンク分析

- 聴覚の周波数分解能は、低い周波数では高く，高い周波数では低い。
- 周波数が高くなるほど幅が広くなる三角窓が並んだフィルタバンクで処理。



energy in each band: $\boldsymbol{M} = m_1, \ldots, m_P$

$$c_i = \sqrt{\frac{2}{P}} \sum_{j=1}^{P} \log m_j \cos\left(\frac{\pi i}{P}(j-0.5)\right), \quad i = 1, \ldots, Q.$$

Mel-Frequency Cepstrum Coefficient: $\boldsymbol{C} = c_1, \ldots, c_Q$

離散コサイン変換

- フィルタバンク分析によって得られた各チャネルの出力を離散コサイン変換する。
- 観測ベクトルの要素同士が無相関であることを仮定できると、音声生成モデルの記述を単純化することができる。
- 離散コサイン変換は信号が十分に滑らかな変化を持つとき、そのベクトルを無相関化することができる。

fb2mfcc.cの穴埋に注意　～ 数式の構造 ～

フィルタバンク分析によって得られた各チャネルの出力 $m_j (j = 1, \ldots, P)$ の対数を離散コサイン変換（**Discrete Cosine Transform: DCT**）

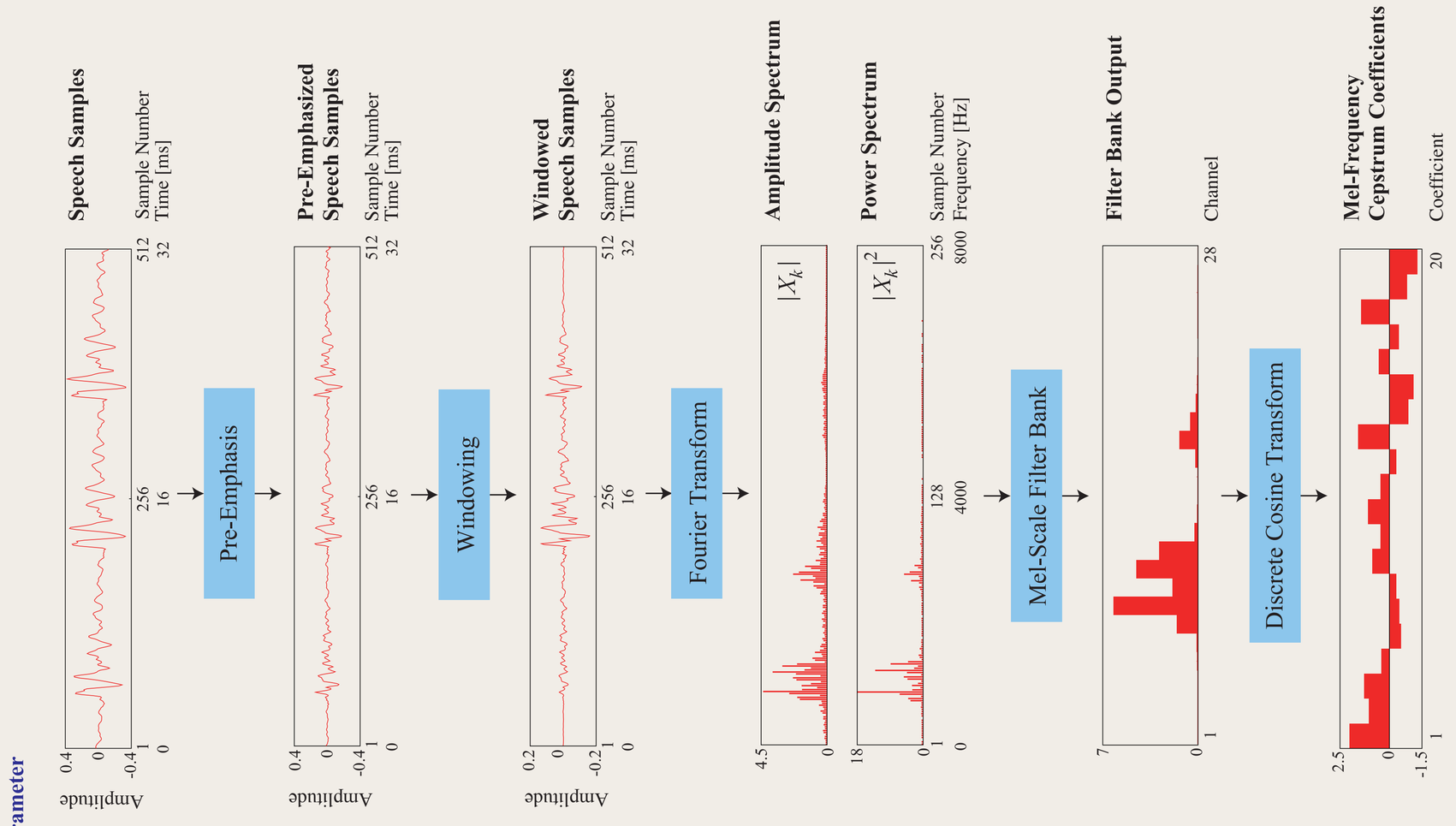
$$c_i \triangleq \sqrt{\frac{2}{P}} \boxed{\sum_{j=1}^{P} \boxed{\log m_j} \boxed{\cos\left(\frac{\pi i}{P}(j - 0.5)\right)}}, \quad i = 1, \ldots, Q \tag{3.4}$$

することにより，メル周波数ケプストラム係数（**Mel-Frequency Cepstrum Coefficient: MFCC**）が得られます．Q はケプストラムの次元数です．

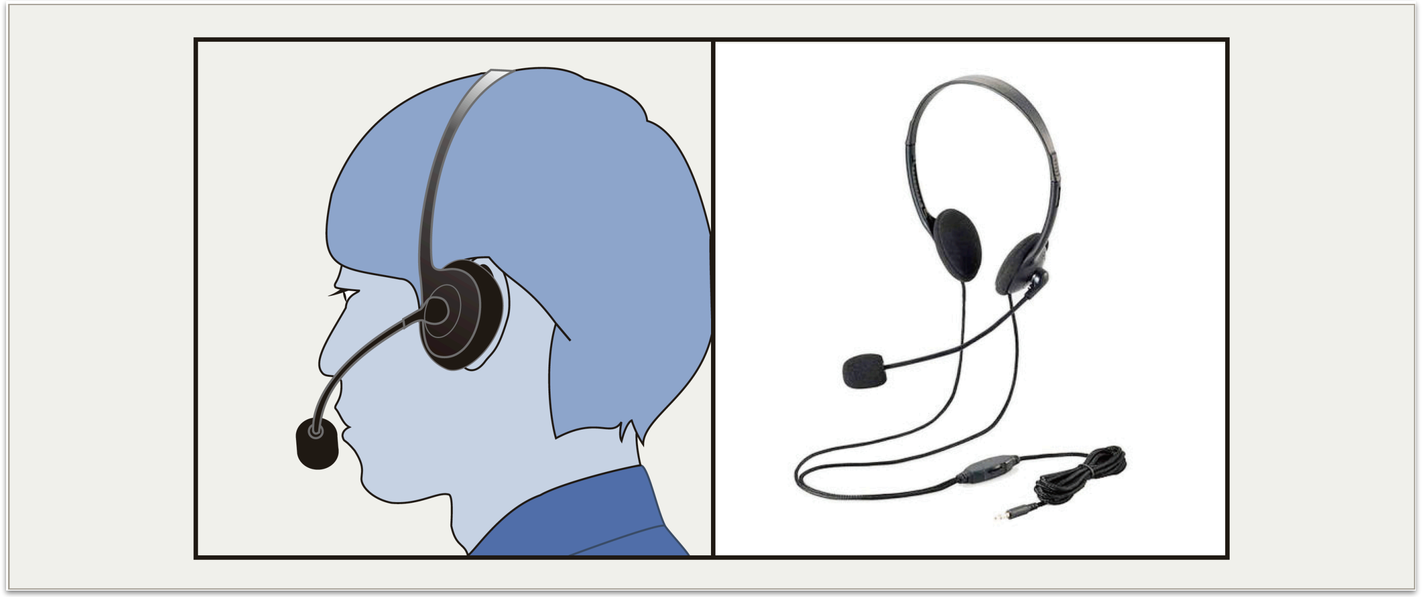
特徴抽出はデータ圧縮

・512点の音声波形 → DFT → 256点スペクトル → Filter Bank → 28次元FB → DCT → 20次元MFCC

・音声認識に必要な情報を抽出しつつ，パラメータ数は 512 から20 まで1／25.6に削減



ヘッドセットの取り扱い



・マイクブームアームの根本の固い部分を持って180 度回転。

・マイクが左側に来るようにヘッドセットを装着。

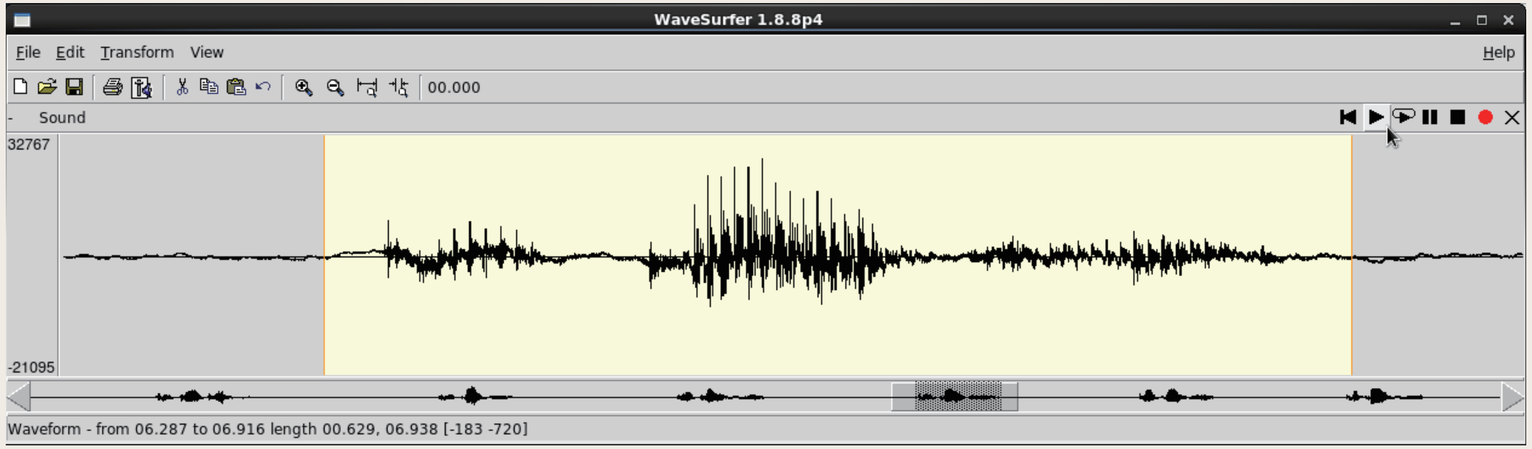
・マイクの角度を変えるときは，本体の取り付け部の固い部分を回す。

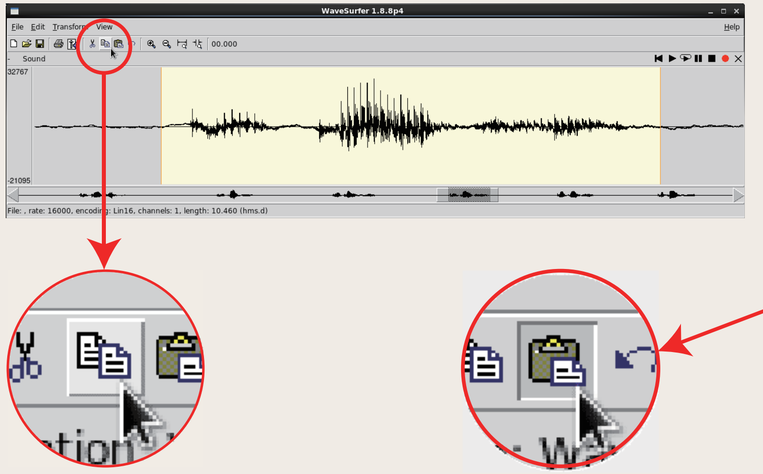
・ヘッドセットを頭に付けずマイクを持って音声入力しないように。

・再生音量を調整するボリュームコントローラーがあるので確認のこと。

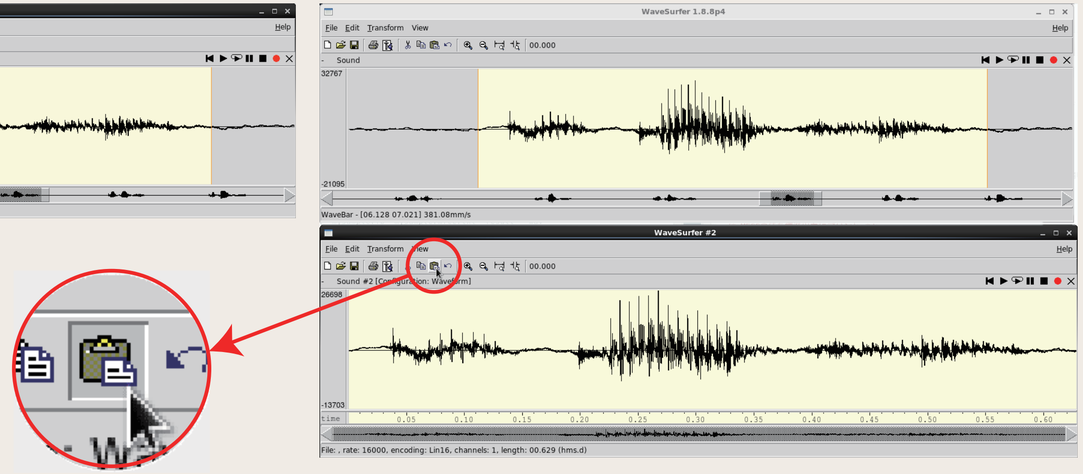
・衛生を考慮し他者と共用しない。

音声の録音、編集、聴取には WaveSurfer を用いる。





(a) 波形を選択してコピーボタンを押す



(b) 新しい窓でペーストボタンを押す

実験第1日のレポート課題1(b)

自分の姓 (family name) または名 (given name) の分析結果

普通に話す。録音データを編集し単語の音声波形部分（＋前後の空白）のみを用いる。

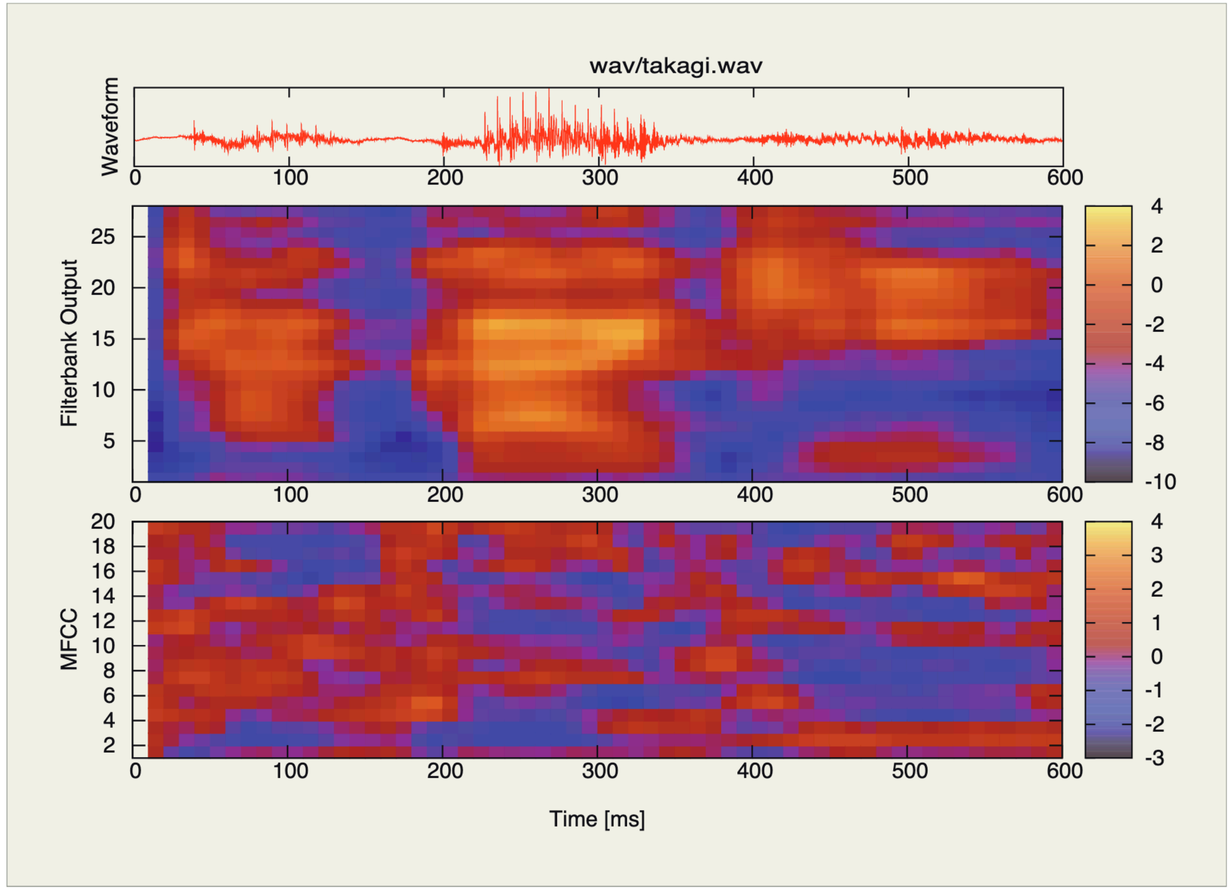


図 3.22: drawspec の出力例．発話は/takagi/．上から，音声波形，フィルタバンク出力，メル周波数ケプストラム係数（MFCC）．横軸は時間 [ms]．

認識タスクの例「名前を言うと電話番号を表示」

単語表記	単語名	発音	音素数	ファイル名
090-1234-5678	ryusei			
090-9876-5432	minato			
080-3571-1131	naomi			
042-443-5001	shuNichi			
090-2018-0523	reN			
03-3456-9999	riN			
011-561-1039	arata			
110	keesatu			
…	…			
…	…			

認識タスクの例「野菜の名前を言うとその野菜の絵文字を表示」

単語表記	単語名	発音	音素数	ファイル名
	hourensou	/hourensou/	9	`hourensou_*`
	jagaimo	/jagaimo/	9	`jagaimo_*`
	kyuuri	/kyuuri/	6	`kyuuri_*`
	nasu	/nasu/	6	`nasu_*`
	ninjiN	/ninjiN/	8	`ninjin_*`
	pimaN	/pimaN/	7	`piman_*`
	tamanegi	/tamanegi/	10	`tamanegi_*`
	tomato	/tomato/	8	`tomato_*`
	toumorokoshi	/toumorokoshi/	13	`toumorokoshi_*`

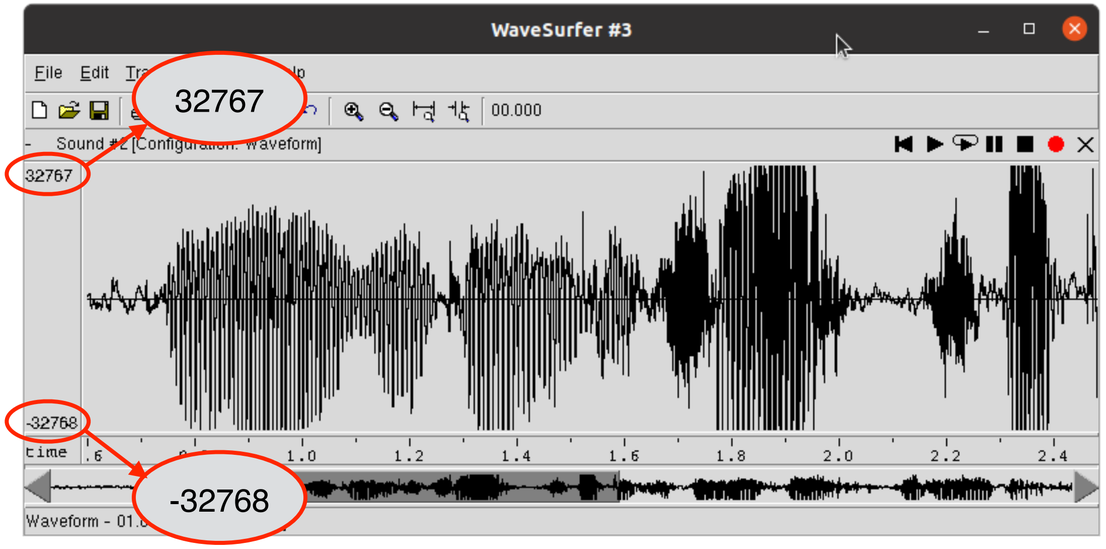
単語音声認識は音声パターンから記号への写像

$$\xrightarrow[\text{パターン認識}]{f} \quad 1, 2, 3, 4, 5, \ldots \text{（記号：単語番号）}$$

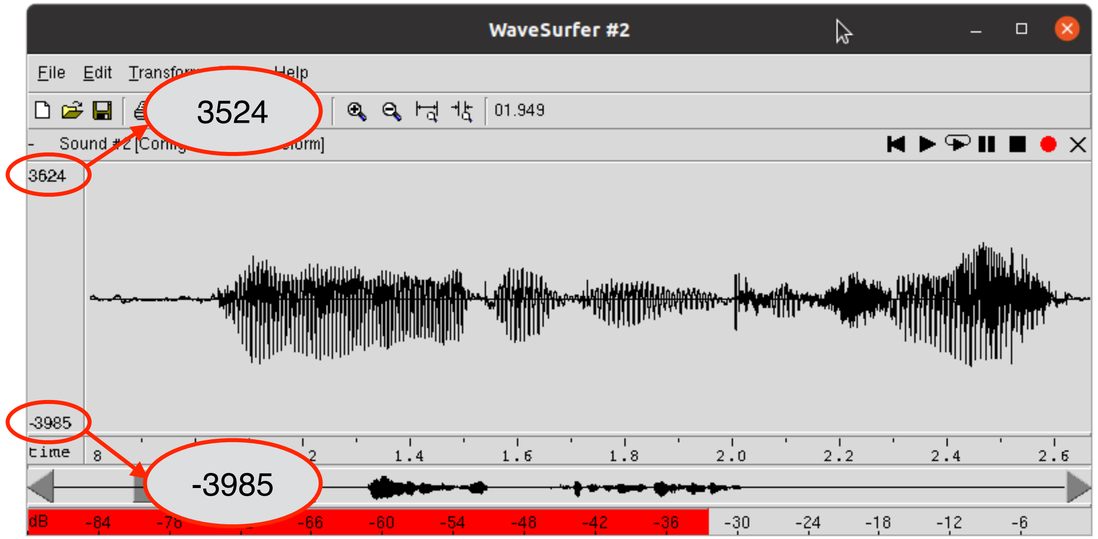
$$\xrightarrow[\text{アプリケーション用後処理}]{} \quad \text{文字列（単語表記など）}$$

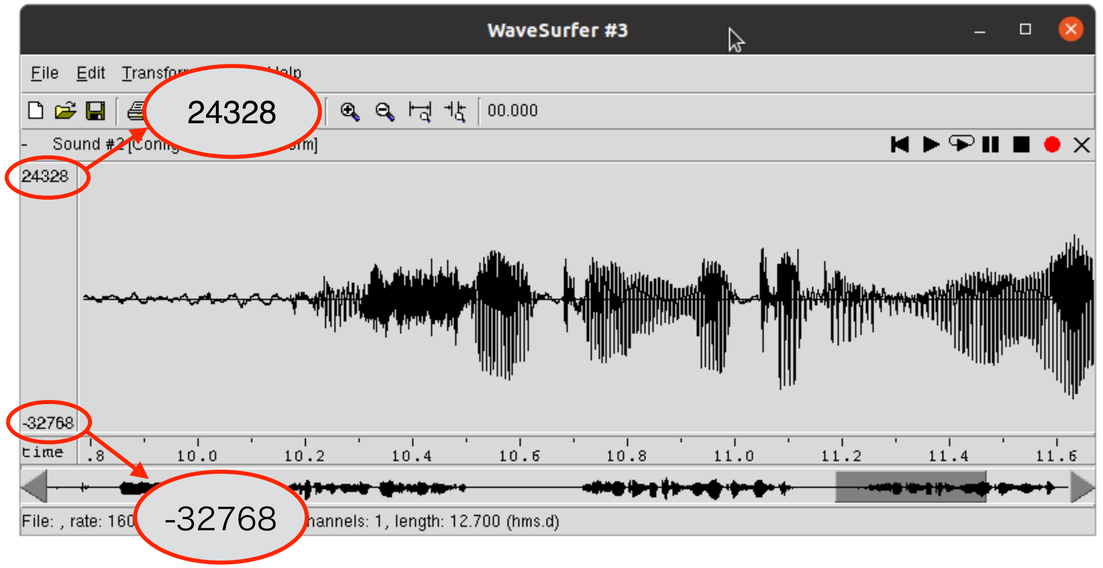
- 個性や創造性を発揮してtaskを考えてみてください。
- 単語の種類は５以上，～７が適当です（作業時間の制約）
- 極端に短い単語や長い単語，　促音「っ」を含む単語があると，この実習の On-The-Fly 音声認識では失敗する可能性が高いので、そのような単語は避けてください.

音声入力の適切な音量…過大になりがち。過小もダメ。



振幅過大で値域を超えた信号は正しく処理できない





音量が適切ならば振幅値域内で
大きなダイナミックレンジを確保できる

振幅過小では波形情報を十分に活かし切れずSN比が悪くなりやすい（量子化誤差が大）

エディタについての注意 ⚠️

Ubuntu付属の「テキストエディタ」は、

本実習の単語認識実験に用いる設定ファイルの

作成に適していないので,

これ以外の開発環境やテキストエディタを使ってください.

レポート

✓ 〆切：実験実施翌週授業日の水曜日23:59:59

2024年11月20日、12月4日、12月18日

✓ 提出先：asr@takagi.inf.uec.ac.jp

- ファイル名は学籍番号_実習回番号（1, 2, 3）_姓.pdfとする。

  （例）2210987_1_Takagi.pdf
- メールの件名はファイル名と同一文字列。
- メール本文には何も書かないこと。

✓ 形式：Portable Document Format（PDF）限定

本日の実習

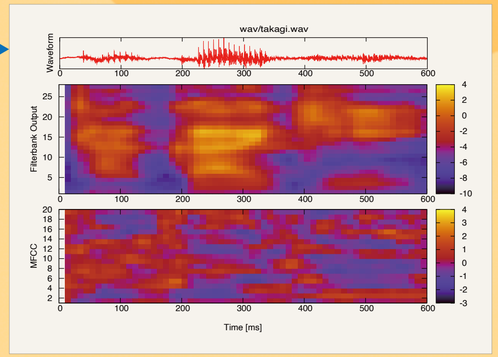
§1.7
計算機実習の準備
ファイルのコピー ＆ 環境設定の追加

§3.10.1
[実習] フィルタバンク分析

§3.10.2
[実習] MFCC 分析（離散コサイン変換）

§3.10.3 -1
[実習] 音声収録の練習

§3.10.3 -2
[実習] スペクトル分析



音声波形

フィルタバンク出力

MFCC

図 3.22　自分の名前（姓または名）のスペクトル

本日時間が余ったら…

単語音声データ（3日目）の録音・保存を推奨

- 認識対象単語が決まっていれば
- 手順はテキスト第6章参照

【理由】

単語音声の学習データ作成作業において、

ラベル付け作業がかなりの労力を要す（約90分）。

録音だけは早めに行っておいて損はない。