2　音声インタフェースを備えたロボットの製作 ～音声インタフェースの構築～

1　本実験の目的

本研究では，音声インタフェースを備えたロボット，即ち，ユーザーの音声コマンドを認識して無線で送信し，それに基づいて動くロボットの製作を通して，信号処理，C プログラミング，無線通信，マイコンプログラミング，ロボット（ハードウェア）製作など，様々な技能の習得を目的としています。最終的には，ロボットを音声コマンド操作することで障害物競走を行います。また，各自の成果を発表してもらい，プレゼンテーション能力の向上も図ります。本実験課題は大きく 1) 音声インタフェース（空気振動としての音声を文字列，即ち，シンボルへと変換する）の構築と，2) シンボル化されたコマンドを無線で送信・受信し，それに基づいて動くロボットの製作，という 2 段階に分かれますが，本書では前半を扱います。

音声（空気の振動）は一次元の時系列信号として観測されますが，そこにはテキスト（文字面）の情報，話者／性別／年齢などの情報，更には感情や意図といった様々な情報が同居しています。一次元の数値列として観測される信号のどこに，これだけの豊富な情報が存在しているのでしょうか？少なくとも人間は，通常誰に教わることも無く，これらを正確に抽出する術を獲得します。本実験前半の目的は，音響信号としての音声を題材とし，下記課題を通して信号処理プログラムについての素養を高め，簡単な孤立単語（ロボット用音声コマンド）認識システムを構築することにあります。

- 音声信号の周期性検出と感情／意図情報の抽出
- 音声信号のスペクトル解析と音韻情報の抽出
- 音声信号のスペクトル解析と話者情報の抽出
- 孤立単語音声（ロボット用音声コマンド）認識システムの構築

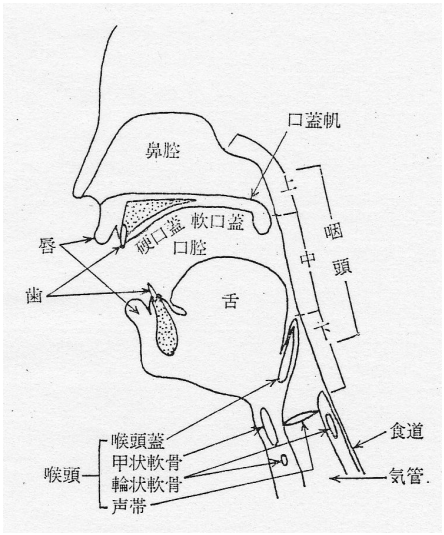
最後に面白い話題提供（問いかけ）をします。半日あれこれ思考してみるのも，時には楽しいものです。

2　音声のいろは　～その生成メカニズムと代表的な音響パラメータ～

2.1　人が生まれながらにして所有する高性能な形状可変楽器＝口

まず「音声がどのように生成されるのか」について簡単に説明します。**図 1** に音声器官を示します。この図を通して母音の生成について考えます。「あー」と発声しながら喉を触ってみましょう。ブルブル震えているのが分かると思います。声帯が振動しています。1 秒間に男性だと 100 回ほど，女性だと 200 回ほど，振動しています。肺によって生成された呼気流が声帯（弁）を通ると，弁の開閉運動が生じ，この開閉運動が振動となって音（音源）が生まれます。しかし，この振動音は「ブー」というブザーのような音でしかありません。まだ「あ」にも「え」にもなっていない，「ブー」といった音でしかありません。しかし，この「ブー」が，口腔（声道）を通って口から放射されると「あ」「い」「う」「え」「お」に化けることになります。即ち，どの母音も音源は同じですが，その後の口の構え，舌の位置などよって音色が変化し，その結果「あ」～「お」になります。言い換えれば，口の構え次第で色んな母音に化けることになります。

管楽器を吹く時に使う（リップ）リードを知っていますか？リード（リップリードの場合は，唇がリードになる）を使って「ブー」を生成し，それを音源として鳴らす楽器が，管楽器です。楽器は違えども，リー



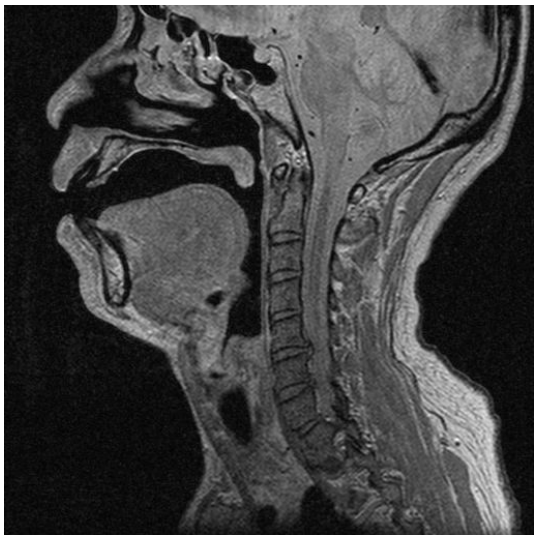


図 1: 音声の生成器官（写真は ATR 提供）

ドで生成される音源はどれも似た音となります。でもそれが，その後の管の形状の違いによって，個々の管楽器特有の音色の違いを生むことになります。「あ」〜「お」の違いが生まれるのも，物理的には同じ原理です。音源として「ブー」を入れ，異なる管形状を持つ口／楽器が「異なる特性を持ったフィルター」として機能し，その出力信号が母音・楽器音ということです。

日本語には 5 種類の母音があります。「あ」〜「お」を発声してみれば分かると思いますが，舌の位置，顎の開け方，更には唇のすぼませ方など，声の通り道（声道）の形状は様々に変化させることができます。「あ」から「い」に連続的に（ゆっくり）変えていくこともできます。言い換えれば，口は無限種類の楽器を構成することができます。日本語というのは控えめな言語で，その中から，たった 5 つの母音楽器しか要求しません。これが英語になると 11 種類となり（但し，単母音のみ），スウェーデン語になると 17 種類となります。少しは異国の地の母音楽器に想いを巡らせておくと，大学院に行ってから苦労しないでしょう。

2.2 母音の生成＝お芋（定常波）の生成

母音生成に関係する発声器官について紹介しました。次に，母音生成についての音響的側面について説明します。高校の物理の時間に「定常波と共鳴」という現象を学びました。同一の進行波と逆行波が重ね合わさって，その場に定常的に居座ったような波として観測される現象です[1]。音波の場合，気柱（音響管）にできる定常波について学んだ学生も多いでしょう。お芋が一つ，二つ，と数えた「あれ」です。その時に，固有振動数（共振・共鳴周波数）についても学んだと思います。下に示す関数を覚えている学生もいるでしょう。気柱の片方を開放端，もう片方を固定端とした場合，固有振動数 F_n は管の長さの関数となります。

$$\text{固有振動数（共振・共鳴周波数）}\ F_n = \frac{c}{4l}(2n-1) \qquad n = 1, 2, 3...$$

c が音速，l が管の長さです。即ち，どのような振動数の波でも居座れる訳ではなく，管の長さが規定する幾つかの振動数（周波数）の波だけが居座れることになります。そして管の形状が，断面積一定の気柱とは異なり，複雑になってくると（例えば「あー」と発声した時の喉の形状を考えてみて下さい），共鳴周波数は管の形状の関数として表現されることになります。ある周波数 f が共鳴周波数となる，ということは，エネルギーが周波数 f 付近に集まることを意味します。結局，管の形状の変化は，エネルギーの分布の様子を変えることとなり，その結果，音色が変わってくることになります。管形状の違いによって，トロンボー

[1] 英語だと，standing waves と言います。こちらの方が理解しやすいですね。

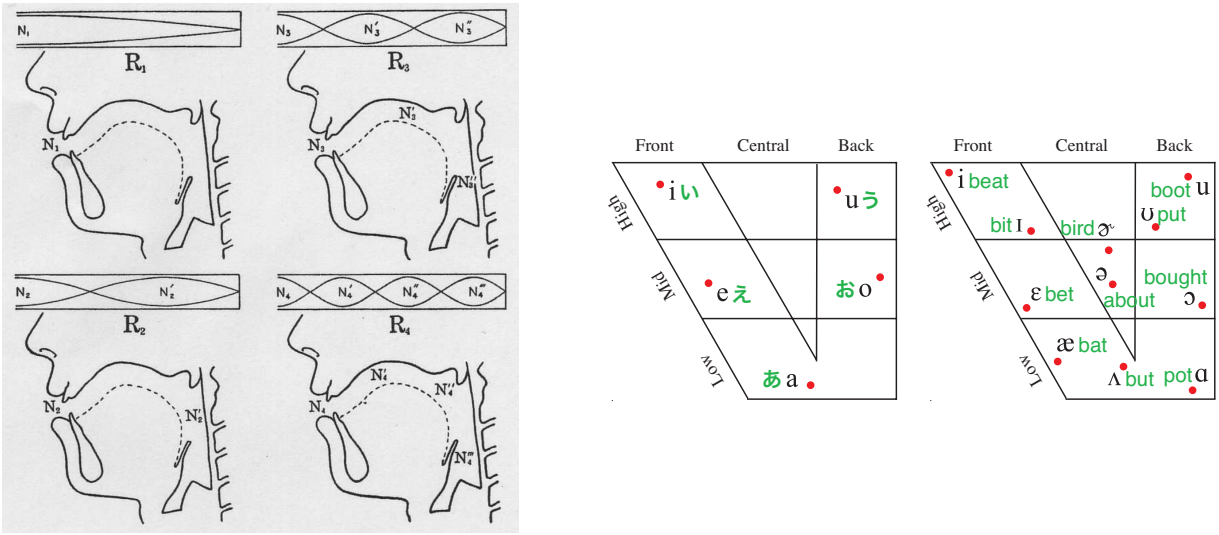


図 2: 断面積一定の声道による弱母音の発声と定常波（第 1〜第 4 フォルマント）／日英の母音図

ンがトランペットにもなり,「あ」が「え」にもなる理由がここにあります。なお，共鳴周波数のことを音声工学では，フォルマント（formant）周波数と呼びます。また，声道をおよそ断面積一定にして母音を発声すると，それは英語の弱母音（曖昧母音，schwa）となります。この時のお芋の様子を**図 2** に示します。

第 2.1 節で「口の形／楽器の形がフィルタとして機能します」と書きました。となれば，そのフィルタの伝達特性 $H(\omega)$ や $H(s)$，或は，その対数パワースペクトルの周波数特性が気になってきます。フォルマント周波数は共鳴周波数（極周波数）ですから，対数パワースペクトル特性のピークに相当する周波数，ということになります。さて「数個の極周波数（フォルマント周波数）を用いてパワースペクトルを表現する」ということは，数個の数値でスペクトルを表現することになりますので，効率的なスペクトル表現と言えます。しかしながら（実験を通して確認して欲しいですが），伝達関数が数式として与えられるのではなく，物理観測量として与えられたパワースペクトルからそのフォルマント周波数を数個求める（推定する）というのは，決して易しい作業ではありません。そのため通常音声認識では，フォルマント周波数は求めずに，スペクトルの概形をそのまま用いて音色を表現することが多いです（ケプストラムの定義を参照)。

2.3　高い声と低い声　〜音声の基本周期と基本周波数〜

「あ」〜「え」がどのように生成されるのか，及び，その違いを音響的に説明しました。さて，高い「あ」と低い「あ」は何が違うのでしょう？どちらも同じ「あ」である以上，フォルマント周波数は凡そ同じです。高い「あ」も低い「あ」も，口の構えは同じですよね。何が違うのか？それは喉を触ると分かります。声帯の振動回数が違ってきます。声帯振動の回数が増加すると我々は「高い」と感覚し，減少すると「低い」と感覚します。即ち，女性の「あ」は声帯振動数が多く，男性の「あ」は声帯振動数が少なくなります。だから，男性の声を「低い」と感覚します。では，何故男性の声帯振動数は少なくなるのでしょうか？これは単に，男性の声帯が「重たい」ために，女性ほど素早く動けないだけ，です。重たいものは動きが鈍い。純然たる力学法則に則って，男性の声は低く，女性の声は高く感覚される訳です。

声の「高い／低い」のメカニズムについては上記の通りですが，音響的にはどのような形で観測されるのでしょうか？声帯振動が早いということは，母音音声の周期が短くなることを意味します（**図 5** 参照)。この周期のことを基本周期と呼び，その逆数を基本周波数と呼びます。ピッチという言葉をよく耳にするでしょう。音の高さについての「心理量」を表す言葉です。そのピッチにおよそ対応する物理現象が，この基本周波数です。心理量／物理量の厳然たる差はあるものの，両者はよく同一視されることがあります。

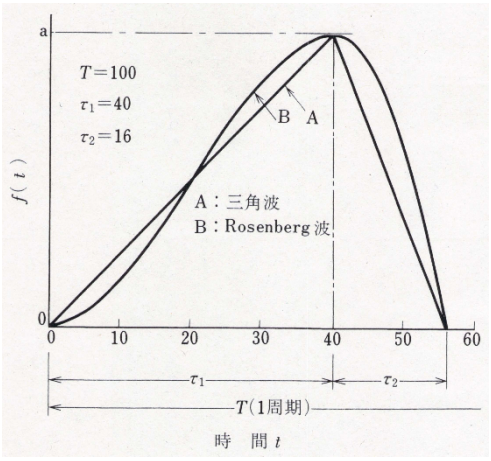


図 3: 三角波としての音源近似

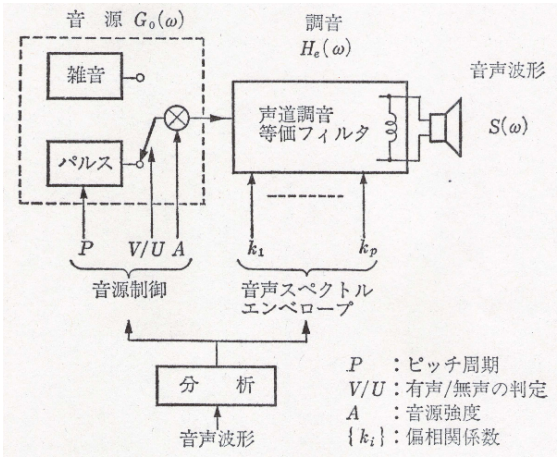


図 4: パルス音源＋声道フィルタ＝母音音声

2.4 パルス音源＋声道フィルター＝音声

声帯付近で生成される音源信号を $g(t)$，声道（フィルター）のインパルス応答を $h(t)$ とすると，出力信号 $s(t)$ は $g(t)$，$h(t)$ の畳み込み積分で得られます。即ち，時間領域では

$$s(t) = g(t) \otimes h(t)$$

となり，これを周波数領域で表現すると，下式が得られます。

$$S(\omega) = G(\omega) \times H(\omega)$$

第 2.1 節で示したように，声道の形状が $H(\omega)$ に相当し，この情報を用いて音声を（例えば平仮名列として）認識することになります。しかし，実際に観測されるのは $H(\omega)$ や $h(t)$ ではなく，$S(\omega)$ や $s(t)$ です。

さて，声帯で生成される「ブー」$g(t)$ はどのような形状をしているのでしょう？これに $h(t)$ が畳み込まれて $s(t)$ が得られます。実は $g(t)$ は（やや滑らかになった）三角波でよく近似されます（**図 3** 参照）。つまり，三角波を声道フィルタに通して出力されたのが母音音声ということになります。音声の生理学的な生成メカニズムとしては「三角波＋フィルタ」となりますが，工学的にはもう少し簡便なモデル化が可能です。三角波は，パルス波をフィルタに通すことでも得られます。そこで，母音音声生成を「純粋なパルス波」に対するフィルタ出力と解釈することも可能です。即ち，パルス波を三角波とするフィルタも $H(\omega)$ に含めて考えよう，ということです。**図 4** では，パルス列と白色雑音を適宜切り替える形で音源をモデル化しています。

観測される $H(\omega)$ ではなく，声道フィルタ特性である $H(\omega)$ の推定ですが，**第 3.5 節**までお待ち下さい。

3 実験課題

3.1 準備 1：音声分析ソフトを触ってみる・使ってみる

まずは，無償でダウンロードできる音声分析ソフトを用いて，自らの声を録音し，波形の描画，スペクトルの計算・描画，フォルマント周波数の計算・描画，ピッチの計算・描画などをさせてみましょう。**第 2 節**で学んだことを，音声分析ソフトを通して再確認し，以降の課題で，各自が，自作のプログラムで抽出することを試みる音響パラメータについても確認しましょう。なお，音声分析ソフトとしては，wavesurfer[2] が良いでしょう。使用法についてまとめた web[3] などもあります。

[2] http://www.speech.kth.se/wavesurfer/
[3] http://www.f.waseda.jp/kikuchi/tips/wavesurfer.html

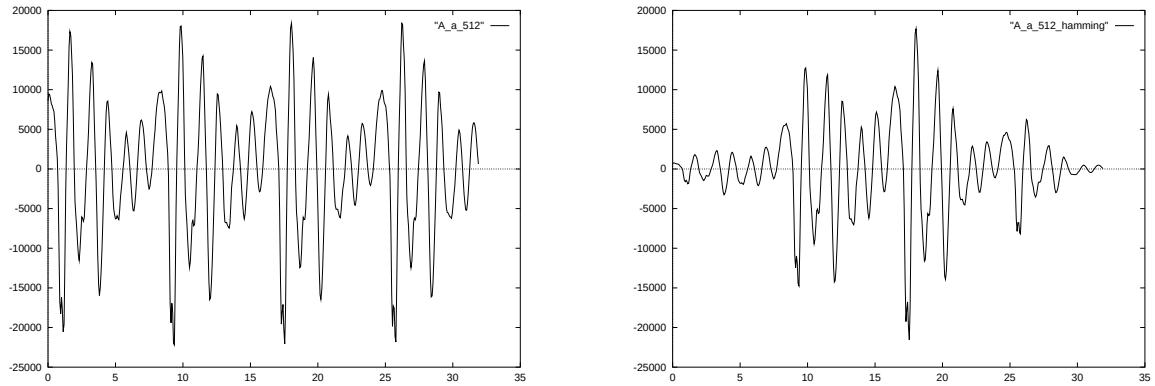


図 5: A_a.XX の先頭 512 個のデータとその窓掛けデータ

3.2 準備 2：音声波形データの読み出しと GNUPLOT による表示

本実験用の音声サンプル[4]を波形表示してみましょう。音声サンプルは全て，16bit/16kHz サンプリングされた音声データであり，short integer の形で格納されています（バイナリーデータファイル）。話者 A の母音/あ/の音声サンプルを先頭から 512 個だけ，GNUPLOT を用いて波形として表示してみましょう。次に，窓長 512 のハミング窓をかけ，再度波形表示してみましょう。なお，ハミング窓の定義は下記の通りです。窓長を N，時刻 $n = 0, 1, 2, .., N-1$ とすると，

$$W_H(n) = 0.54 - 0.46 \cos\left(\frac{2n\pi}{N-1}\right)$$

となります。以降で行うピッチ抽出やスペクトル包絡推定は，窓長は 512 に固定し，何れの場合もハミング窓をかけた波形に対して行います。GNUPLOT による表示ですが，**図 5** のような図が得られるはずです。

学生が使用する計算機によって CPU のエンディアンが異なるため，big/little の両エンディアンのデータを用意しています。必要なものをダウンロードして下さい（wav 化したものも置いています）。

3.3 課題 1：自己相関関数を用いたピッチ抽出

基本周波数（ピッチ）は，音声波形の周期に対する逆数で定義されます。周期を求める常套手段として，信号の（正規化）自己相関関数を求め，相関値のピークを示す遅れ時間幅 τ を計算する方法があります。

自己相関関数を

$$r_\tau = \sum_{t=0}^{N-1-\tau} s_t s_{t+\tau}$$

とすると，正規化自己相関関数は，

$$v_\tau = \frac{r_\tau}{r_0}$$

となります。用意された音声サンプルの幾つかについて，先頭 512 個のデータに対する正規化自己相関関数を求め，ピッチの抽出を行いましょう。そして，前節で表示した母音波形から視察により基本周期を調べ，自動抽出した結果の正当性を検討しましょう。抽出エラーがある場合，どのような原因が考えられるのか，考察しましょう。なお，正規化自己相関関数の図としては**図 6** のような波形が得られるはずです。

次に，幅 512 点の窓を 256 点ずらして次の窓として（シフト長＝窓長/2）再度基本周波数を求めてみましょう。窓を次々とずらし，音声ファイル全体に渡って基本周波数を時系列として求めてみましょう。その

[4] http://www.gavo.t.u-tokyo.ac.jp/~mine/jikken/speech_interface

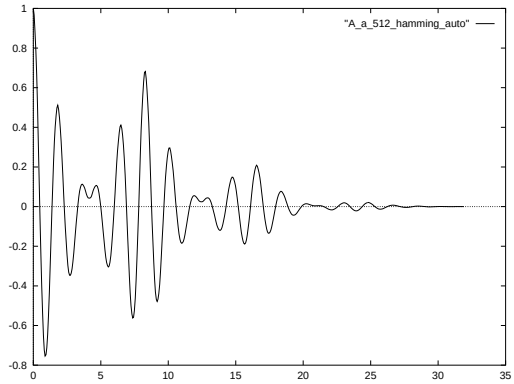


図 6: A_a.XX の先頭 512 個から得られた正規化自己相関関数

結果を「横軸を時間（シフト回数），縦軸を log(基本周波数)」として図示しなさい。各自「あいうえお」を様々なイントネーションパターンで発声したものをファイル化し，それに対してピッチ抽出を試みてみましょう。意図したパターンが得られたのか，よく検証して下さい[5]。

自己相関値のピーク検出は，音声波形そのものよりも，音声波形をローパスフィルタ (cutoff=800Hz ほど）に通した波形から求めた方が，その精度が高くなることが知られています。余力のある者はローパスフィルタを構成して，実際に検討してみるのも良いでしょう。

3.4 課題 2：DFT/FFT を用いたスペクトルの算出と描画

DFT（離散フーリエ変換）を用いて音声信号を周波数解析し，スペクトルとその包絡特性を抽出してみましょう。DFT は次式で得られます。時間軸上の複素時系列 $(x_r(n), x_i(n)), n = 0, ..., N-1$ の DFT は，

$$\begin{aligned} X_r(k) &= \sum_{n=0}^{N-1} x_r(n)\cos(2\pi nk/N) + x_i(n)\sin(2\pi nk/N) \\ X_i(k) &= \sum_{n=0}^{N-1} -x_r(n)\sin(2\pi nk/N) + x_i(n)\cos(2\pi nk/N) \end{aligned}$$

であり，逆 DFT は，

$$\begin{aligned} x_r(n) &= \frac{1}{N}\sum_{k=0}^{N-1} X_r(k)\cos(2\pi nk/N) - X_i(k)\sin(2\pi nk/N) \\ x_i(n) &= \frac{1}{N}\sum_{k=0}^{N-1} X_r(k)\sin(2\pi nk/N) + X_i(k)\cos(2\pi nk/N) \end{aligned}$$

となります。なお，音声のように実数時系列を DFT にかける場合は，$x_i(n) \equiv 0.0$ とすればよい。得られた $X_r(k)$，$X_i(r)$ に対して，対数パワースペクトル（ピリオドグラム）は下記式で得られます。

$$P(k) = \log_{10}\left[\frac{1}{N}\left\{X_r(k)^2 + X_i(k)^2\right\}\right]$$

用意された音声サンプルの幾つかについて，先頭 512 個のデータに対する対数パワースペクトルを求め，その周波数特性を描画してみましょう。**図 7** のような図が得られるはずです。

DFT に対する高速演算として FFT があります。実験 web[6]に FFT の穴埋めソースプログラムを用意しています。こちらをダウンロードし，適切に穴を埋め，同様の作業を FFT を用いて実行しなさい。DFT と FFT の計算速度（計算量）の比較を行い（理論値），実験値と比べてみましょう。

[5]「雨」と「飴」などの同音異義語を収録してピッチ抽出するのも良いだろう。

[6]http://www.gavo.t.u-tokyo.ac.jp/~mine/jikken/speech_interface

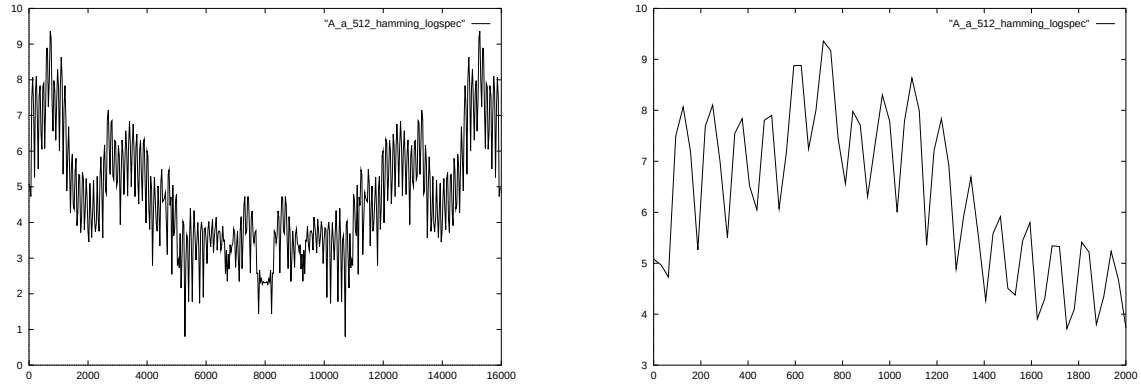


図 7: A_a.XX の先頭 512 個から得られた対数パワースペクトルとその拡大図

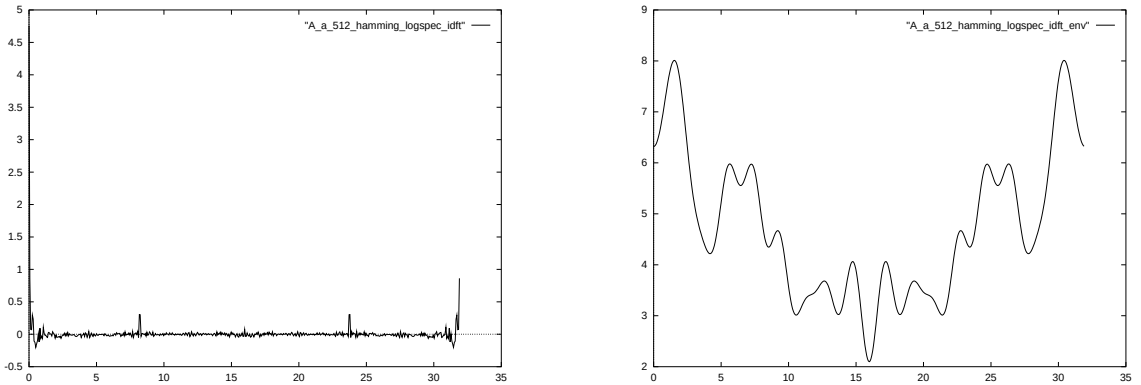


図 8: A_a.XX の先頭 512 個から得られたケプストラム係数とスペクトル包絡

3.5 課題 3：FFT を用いたケプストラム係数の計算とスペクトル包絡の推定

第 3.4 節で求めた対数パワースペクトルは，$S(\omega)$ に相当し，$H(\omega)$ ではありません。$S(\omega)$ から $G(\omega)$ の影響を除去し，$H(\omega)$ を求める方法として，ケプストラム分析法が知られています。

$$S(\omega) = G(\omega) \times H(\omega)$$

に対して，対数パワースペクトルをとると，

$$\log_{10}|S(\omega)| = \log_{10}|G(\omega)| + \log_{10}|H(\omega)|$$

と，$G(\omega)$ の影響は加算成分となります。パルス列のパワースペクトルもパルス列となります。この時の周波数軸上の周期は，元のパルス列の周期の逆数（基本周波数）となります。

結局，時間領域のパルス列間隔が基本周期，周波数領域のパルス列間隔が基本周波数となります。即ち**図 7** において，細かい（櫛状の）ピークは基本周波数間隔で並んでいます。これが声帯振動によるスペクトルへの影響であり，これを除去（平滑化）することで $\log_{10}|H(\omega)|$ が推定できます。

FFT を用いて櫛状のピークを除去します。**図 7** に対して逆 FFT を行い，得られた係数（これをケプストラム（cepstrum）係数と呼ぶ）の高域を全てゼロにして FFT をして元に戻します。こうすると櫛状のピークは除去されます（FFT を使って，ローパスフィルタを構成しています。**図 8** 参照）。ケプストラム係数の単位はケフレンシー（quefrency）と呼ばれますがが，これは物理的には，時間と同義です。

複数話者の 5 母音の孤立発声サンプルに対して，各々先頭の 512 個のデータ（+窓掛け）からケプストラムを抽出し，各話者，各母音のスペクトル包絡を描画してみましょう。母音毎にスペクトル包絡が異なるこ

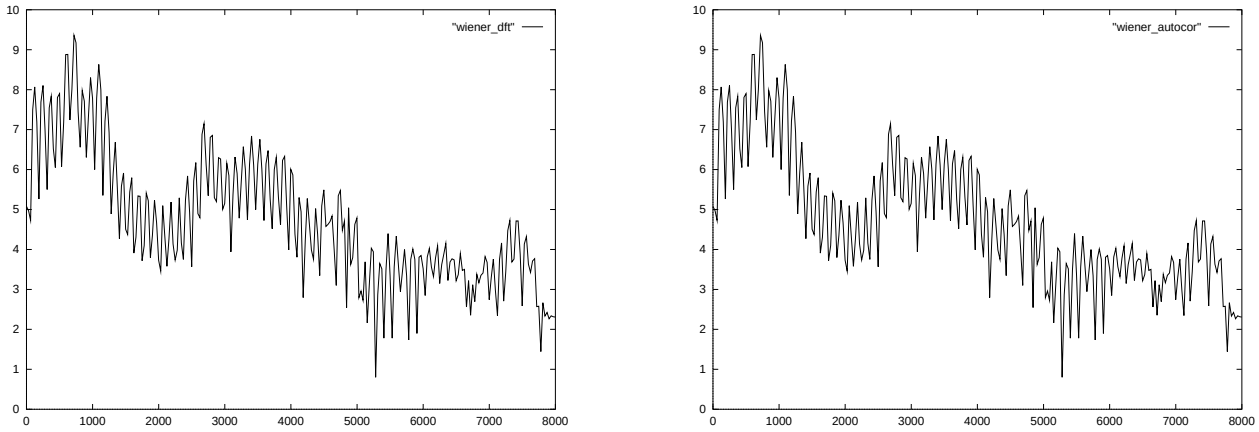


図 9: FFT による対数パワースペクトルと自己相関より求めた対数パワースペクトル

と，話者が異なっても母音が同じであれば，およそスペクトル包絡は同じであることなど，音声の物理的特性を検証してみましょう（スペクトルを比較する場合は，フォルマント周波数を比較すればよい）。なお実数系列の FFT の結果は，窓中心に対して，共役複素数の関係になります。当然のことながら，対数パワースペクトル化した場合は，窓中心（ナイキスト周波数）を中心として実数の対称形となります。この対称形をそのまま逆 FFT （ケプストラム化）すると，再度実数の対称形となります。ケプストラムの高域をカットする場合にこの対称形を崩すと，対数パワースペクトルが対称形とならず，かつ，虚数項を持つようになってきてしまいます。高域カットの時には，対称性を崩さないように注意しましょう。

抽出したケプストラムにおいて，中ケフレンシー領域に小さなピークが見つかるはずです。高域カットの後に，このピークが残る場合と残らない場合とで，平滑化パワースペクトルはどのように変化するのか検討してみましょう。また，このピークは音声の何に対応しているか考察しましょう。なお，音声認識で使われるケプストラムパラメータは 0 次元〜十数次元までの場合が殆んどです。

3.6 課題 4：パワースペクトルと自己相関関数（Wiener-Khintchine の定理）

以上で，音声信号の自己相関関数，及び，FFT に基づくパワースペクトルが算出できたことになります。自己相関関数とパワースペクトルの間には，有名な Wiener-Khintchine の定理と呼ばれる関係があります[7]。この定理の正当性を実際にデータを通して検証しなさい。ここで，以下の点に注意しなさい。窓掛けの有無，自己相関関数の正規化の有無（正規化時の分母係数），FFT による演算が仮定する時系列信号の特性などを適切に考慮しないと，次のような結果は得られないはずです。**図 9** は，FFT による対数パワースペクトルと自己相関関数から得られる対数パワースペクトルです。計算誤差を無視すれば同一になります。

3.7 準備 3：ケプストラム係数のユークリッド距離に対する物理的意味について

声道形状の違いが音色（周波数軸に対するエネルギー分布＝スペクトル包絡）の違いを生むことは既に説明した通りです。言い換えれば，与えられた音の「あ」らしさ，や,「い」らしさは，スペクトル包絡特性に表れることになります。ここでは，音声の二つの対数パワースペクトル包絡が与えられた時に，その包絡間の差異（距離）を計算することを考えます。二つの対数パワースペクトル包絡曲線を $f(\omega)$, $g(\omega)$ とします。なお，両包絡ともケプストラム係数を $0 \sim N$ 次元まで使った包絡曲線とします。この時，対数スペクトル包絡の平均値を除去したスペクトル包絡（つまり，$1 \sim N$ 次元のケプストラム係数による包絡）の差

$$D(\omega) = \left[f(\omega) - \overline{f(\omega)} \right] - \left[g(\omega) - \overline{g(\omega)} \right]$$

[7] え！この定理を知らない？そういう学生は，各自調査しておきなさい。

の 2 乗の全積分値に対して，以下の関係が成立します。

$$\frac{1}{(2\pi)^2}\int_0^{\pi} D(\omega)^2 d\omega = \sum_{i=1}^{N}\{c_f(i)-c_g(i)\}^2$$

$c_f(i), c_g(i)$ は，各々 f, g に対する i 次ケプストラム係数です。このように,「ケプストラム係数（1 次元〜）のユークリッド距離（の 2 乗)」は「両スペクトルの包絡間差異（の全積分値)」という物理的に明確な対応物があるため，二音の音色差を定量化する場合に，頻繁に用いられます。なお，FFT の定義から明らかですが，ケプストラムの 0 次項はパワーの情報を持つため，音声認識では用いられないことの方が多いです[8]。

3.8 課題 5：スペクトル包絡を用いた話者の同定

第 2.2 節で，フォルマント周波数は声道の長さ／形状に依存することを数式として導きました。舌の位置を変えれば形状変化が起こり，フォルマント周波数は変化します（「あ」〜「お」の違い)。しかし，形状変化は話者の違いによっても生じます。例えば，子供と大人は声道の長さが異なるため，フォルマント周波数は大人だと下がり，子供だと上がることは，定常波の固有振動数を求める式からも容易に想像できます。言い換えれば，スペクトルの包絡特性は，音韻情報のみならず，話者情報も多分に含んでいます[9]。

第 3.5 節では，母音音声スペクトルに観測される音源情報（櫛状のスペクトルピーク列）を除去するために，スムージングを施しました。話者情報を抽出するためには，母音の違いがもたらすスペクトル包絡への影響を除去して，話者の違いがもたらすスペクトル包絡への影響のみを抽出できればよいことになります。種々の母音が連続して発声された状況を考えましょう（「あえうおえいあお ‥」）この場合，母音が変化する度にスペクトルは動きます。さて，様々に動いたスペクトル時系列（ケプストラムベクトル時系列)に対して，時間方向に求めた平均ケプストラムベクトルを考えます。この平均ベクトルは何を表現しているでしょうか？種々の母音が発声されていれば，平均ケプストラムベクトルは,「母音によるスペクトル変化」がキャンセルされ，その音声スペクトルが元来持つ「偏り」成分（「直流」成分）が抽出されたことになります。そしてこの「偏り」成分こそ，その音声の「話者性」と解釈することができます。

ちなみに,「あ」〜「お」の 5 母音を音響空間にプロットした場合，その重心の位置に来る母音は，実は，弱母音（曖昧母音，schwa）になります。また弱母音を生成する時の舌の位置は，口腔内の中央に位置しています。つまり，弱母音は音響的にも，調音的にも（音声生成器官のことを調音器官とも言う）中心に位置する母音となります（**図 2** 参照)。この図には断面積一定の音響管で生成される母音が弱母音であることも示しています。この音響管に対して，狭めを生じさせたり，局所的に太くしたりすると，それが「あ」〜「お」の母音となります。即ち「あえうおえいあお ‥」の直流成分を求める，ということは，口腔内の中心位置に舌を配置し，断面積一定となるように声道形状を整えて生成される母音（弱母音）のスペクトル特性を推定することになります。これを「話者性」として解釈しよう，ということになります。

数名の話者から 10 秒ほどの任意母音連結音声と「あいうえお」という音声を収録しなさい。次に，無音区間が無いことを確認してから音声をケプストラム時系列化し，任意母音連結音声の平均ベクトルを話者別に求めなさい（各話者モデルの構築)。各話者の「あいうえお」音声の各時刻のケプストラムベクトルが,各話者モデルとどれだけ離れているのかを計算し（1〜N 次までのユークリッド距離でよい),「あいうえお」の全区間を通して一番近いと判定された話者を選定してみましょう。この様にして任意の「あいうえお」音声に対して，事前登録された話者の中から，一番近いと思われる話者を推定することができるようになります。各話者の「あいうえお」音声に対して，正しく話者が同定されるかどうかを検証してみましょう。同定精度が悪い場合，精度向上のための工夫を考えましょう。もし，知人に一卵性双生児の兄弟や姉妹がいれば，彼らの声がどれだけ似ているのかを数値化することも可能でしょう[10]。

[8] 大きな声の「あ」と小さな声の「あ」を区別する必要があれば，積極的に使うべきである。
[9] だから，声を聞いて話者を同定することができる。
[10] 結局，彼らの声が似ているのは，彼らが持つ楽器の形状が「うり二つ」なのが原因です。

3.9　課題 6：各自が持参した音サンプルの音響分析

課題 1 にて，イントネーションやメロディーを考える際の主要な音響的特徴である，基本周波数（や基本周期）の抽出方法を学びました。また，**課題 3** や**課題 5** で，音韻性や話者性を考える際の主要な音響的特徴である，対数パワースペクトル包絡の抽出方法を学びました。何れの課題も C 言語で自作プログラムを書き，一次元の時系列信号として観測される音声信号から，所望の情報を抽出したことになります。ここで，再度 wavesurfer を起動して，このツールで何ができるのか，あれこれいじってみて下さい。恐らく**準備 1** の時に比べれば，遙かに「中身の処理」について，想像できるようになっているのでは無いでしょうか？

例えば,「信号処理工学」の授業でも一部扱いましたが，短時間フーリエ解析を行う場合，その窓幅の設定によって，基本周波数（の情報が）どのようにスペクトルに現れるのかは異なってきます[11]。1) wavesurfer の分析条件をあれこれ変えて音響分析を行い，各自の信号処理に関する知識を確認（整理整頓）しましょう。2) 更に，各自が興味ある音（声）サンプルを持ち寄り，音響分析をかけてみましょう。楽器音や動物の鳴き声を音響分析するもよし，物まねタレントの声帯模写音声と，本人の声の違いを分析するもよし，日本人が苦手な right と light の（音響的）違いを分析するもよいでしょう。

なお**課題 7**（次の課題）は，受講学生全員を対象とした事前のレクチャーが必要となります。そこで，**課題 6** は各学生の進捗状況を合わせるためのバッファー的な意味合いを持たせています。進捗が早い学生は，是非ともあれこれ音を持ち寄って，面白い音響分析を行ってみて下さい。

3.10　課題 7：孤立単語音声認識システムの構築

本課題では，孤立単語音声認識システムの構築を目指します。最終的な「音声コマンドを用いたロボット操作」における音声インタフェースの構築です。構築すべきシステムの概要（仕様）を下記にまとめます。

> マイク入力された音声を常時モニターし，発話頭＆発話尾を（つまり，発話区間を）検出する。検出された発話区間に対して，想定した単語セットの中で，どの単語が発声されたと仮定するのが（音響的に）妥当なのかを計算し，最も妥当であると判定された単語を結果として標準出力に出力する。この作業を，ポーズで挟まれた音声区間に対して，延々と繰り返す。

例えばユーザは,「前進　回れ　右　後退」と言った感じで，コマンドをポーズを挟みながら発声します。システムは，各発声（音声信号）を逐一認識し，文字列（シンボル）へと変換します。この文字列をロボットが扱えるコマンドに変換して，無線（Bluetooth）で送信し，ロボットを操作します[12]。

このような音声インタフェースを構築するには，**課題 6** までの知識は不可欠ですが，決して十分ではありません。その他にも様々な知識を獲得し，技術を学ぶことが必要です。本実験課題では，それら全ての習得は目的としません。その代わり，上記の仕様を満たす（最低限の）システムを学生に提供し，そのシステムを目的に応じて修正する形で，各班でオリジナルな音声インターフェイスを構築してもらいます。

以下，提供する孤立単語音声認識システムのモジュール構成（**図 10**），及び各モジュールを説明します。

音響分析モジュール 入力音声を逐一，ケプストラムベクトル時系列へと変換します。また，ケプストラムの 0 次元はパワー成分となりますが，これを参照することで発話頭や発話尾の検出を行います。

音響モデル 音素[13]を単位とした音声の統計的モデル。「あ」という声は話者によって異なりますので，多くの話者から「あ」の音を集めて（即ち，ケプストラムベクトルを大勢の話者から集め），それを多次

[11] **図 9** は 512 点を窓長とする分析です。例えば，窓長を基本周期より短くすると，やがて，櫛状のスペクトルピークは見られなくなります。この場合，基本周波数の情報はどのような形でスペクトルに現れるのでしょうか？

[12] なお，標準出力に出力するのは，認識結果の単語（文字列）だけでなく，その単語の音響的妥当性（数値データ）も出力できる。学生に提供するシステムの仕様変更については，実験を進めながら TA の方で適宜考慮する。

[13] 英語の辞書にあるような，発音記号を考えて下さい。

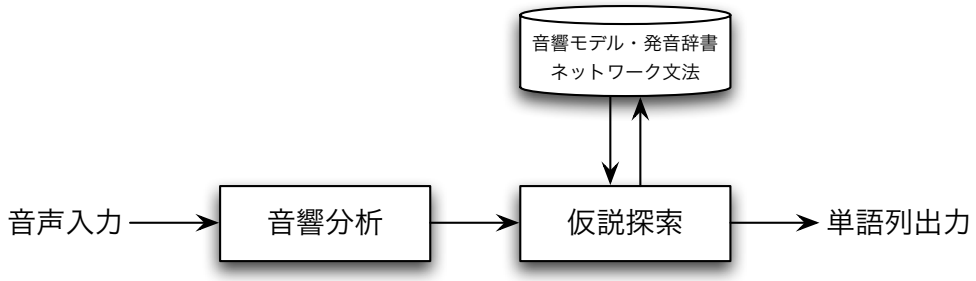


図 10: 孤立単語音声認識システムの概要

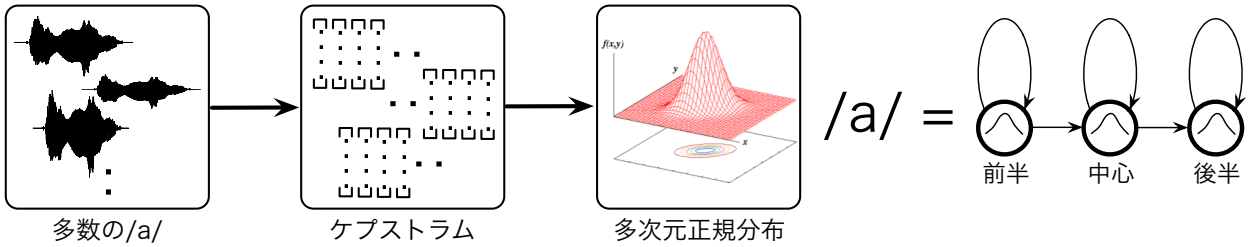


図 11: 多次元正規分布による音響事象のモデル化と音素の音響モデル

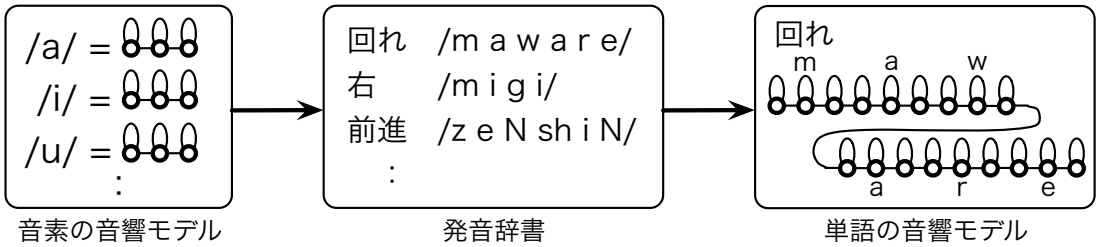


図 12: 発音辞書（音素の音響モデルを用いた単語の音響モデル構成法）

元ガウス分布[14]を用いて近似します。一言で「あ」と言っても音響的には，直前の音素からの遷移部分（前半），安定部分（中心），直後の音素への遷移部分（後半）から構成されると考え,「あ」を，N 次元ガウス分布を 3 つ並べたものとしてモデル化します（**図 11** 参照）。日本語の音素数は約 40 あります。各音素に対して統計的な音響モデルを一つ用意します。

発音辞書 各音素の音響モデルと,「各単語がどのような音素列として記述できるのか」の情報さえあれば，任意単語の統計的な音響モデルが構成できます。即ち，上記の音素モデルを並べることで単語モデルは構成されます。発音辞書は，各単語と音素列との対応表です（**図 12** 参照）。場合によっては，一単語が複数の発音を持つこともあります[15]。その一方で，ケプストラムはピッチの情報を持たないため，アクセント型が異なる「飴」と「雨」は原理的に識別不可能となります。

ネットワーク文法 想定する語彙セットの各単語の統計的音響モデルがあれば，音声認識が出来る訳ではありません。例えば，千単語の辞書を想定し，毎回「1/1000」という難易度の単語同定をすれば，認識率は極めて低くなります。この場合，次にどんな単語が来易いのかを予測しながら（次に来るであろう単語を絞り込みながら）認識することが必要です。最終的には，数十〜200 程度にまで絞らないと現在の技術レベルではまともにシステムは動きません。このような単語制約の一つの実装方法として，ネットワーク文法があります（**図 13** 参照）。これは，一発話中の単語列のバリエーションをネット

[14] 1 次元ガウス分布は平均値と分散で表現されますが，N 次元ガウス分布となると，平均値は N 次元の平均ベクトルに，分散は $N \times N$ の分散共分散行列に置き換わります。集められたケプストラムベクトル群から，これらのパラメータを推定します。

[15] 例えば，永遠＝えいえん，とわ，とこしえ，など。

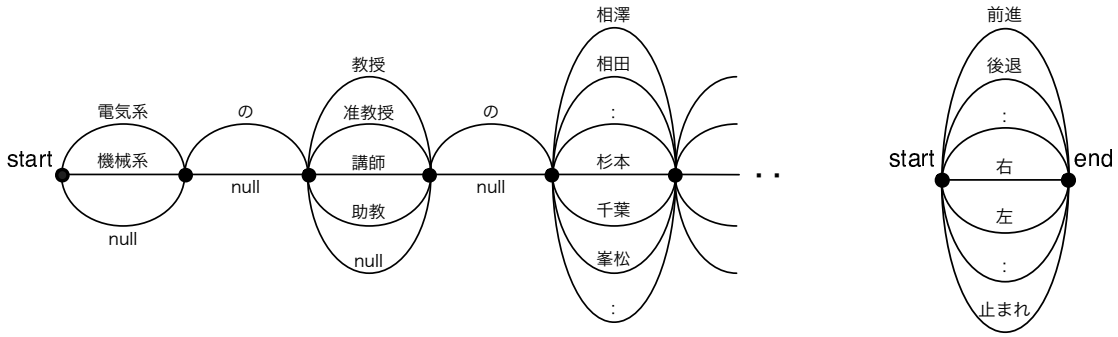


図 13: ネットワーク文法（本実験で使うネットワーク文法は右図のようになる）

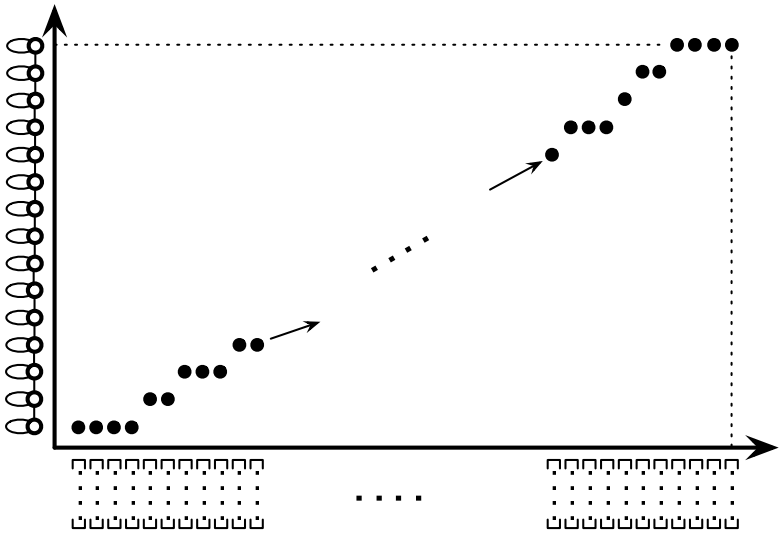


図 14: 単語音響モデルに対する最適パスの探索

ワーク状のグラフで表現し（開始・最終ノードも規定されています），開始ノードからどのノードを辿って最終ノードに到達するパスが音響的に最も妥当なのかを計算します。得られた最適パスに沿った単語列を，発話中の単語列として出力します。なお今回の実験では，一回の発声の中に存在する単語数は 1 ですので，ネットワーク文法は原理的には必要ありませんが，下記に示すような「音声対話システム」を構成する場合はネットワーク文法は必須の技術となりますので，説明を加えました。

仮説探索モジュール（デコーダ） 入力された音声（ケプストラムベクトル時系列）を，ネットワーク文法のどのパスに通すのが最も音響的妥当性が高くなるのかを計算します（仮説探索）。最高スコアのパス上の単語列が認識結果です。**図 14** に，ある単語音響モデル一つに対して行った最適パス探索の様子を示します。単語一つだとネットワーク（グラフ）も何も無いため探索できないように思うかもしれませんが，そうではありません。どこからどこまでの音声区間が（音響モデル中の）何番目の分布に対応するのか，は無数の組み合わせが可能です。その中から，音響的に最も妥当な「単語音響モデルと入力音声間の対応付け」を求めています（即ち，探索です）。さて，**図 13** に示したネットワーク文法において，各単語を「単語音響モデル」で置き換えれば，そのネットワーク文法が受け付ける「文集合」全体を表現する巨大なネットワーク型音響モデルができます。**図 14** で示した最適パス探索は，縦軸に left-to-right 型の単語音響モデルを配置しましたが，実はここに，ネットワーク型の音響モデルを配置することもできます。つまり，ネットワーク文法から構成されるネットワーク型音響モデルに対して，最適パス探索が可能です。そして，得られた最適パス上の単語が認識結果となる訳です。

デコーダ出力の受け取り方 今回用意する雛形システムは，強制終了しない限り，延々とポーズで挟まれた音声区間を一単語として解釈し，認識するシステムとなっています。この認識結果を，外部プログラム（例えばロボットに Bluetooth 通信するプログラム）側で参照する必要があります。外部プログラムによる参照を実装したサンプルプログラムも提供します。

まずは雛形システムに 30 単語ほどを登録し，全単語を一つずつ読み上げて認識実験を行ってみましょう。声は人によって異なりますから，認識率の高い人，低い人がいて当然です[16]。また，外部マイク入力と，内蔵マイク入力で比較するのも良いでしょう。班内で一人一人が個別に認識実験を行う場合と，同時に行う場合（つまり別話者の発声がノイズとして混入する場合）とで比較するのも良いでしょう。ある条件の認識性能が著しく悪い場合，その条件での音声と，良好な条件での音声とを収録し，両者のスペクトルを比較してみて下さい。人間が聞くと，(人間の）音声認識率は劣化しないと思われる音声でも，機械を相手にすると認識率の大きな劣化をもたらすことは，頻繁にあります。何故，機械の音声認識はまだまだ貧弱なのか？逆に言うと，何故，人間はそんな音声でも楽々と扱えるのか？考えてみて下さい。

余裕のある班は，ネットワーク文法を修正して，例えば，受付嬢ロボット（工学部 2 号館 4 階の案内嬢ロボット）に搭載する音声インタフェースを試作するなどしても良いでしょう。受付嬢に投げかける発話を想定して，それを網羅するようなネットワーク文法を構築し，入力音声を認識し，単語（テキスト）列に変換するインタフェースです。発話表現が複雑になると，一気にネットワークが煩雑になって来ませんか？

なお，実際の孤立単語音声認識システムや，その使い方，各種設定ファイルの書き方など，具体的な作業については，実験 web[17]を参照して下さい。時として，学生側の要求を TA が呑んでくれるかもしれません。

本実験課題は，最終的に「音声コマンドを発声することでロボットを操作する」ことを目指します。ですが,「孤立単語音声認識システム」が出力した（文字列化・シンボル化した）単語を，そのまま，ロボットに無線送信する必要はありません。出力単語一つがロボットの動作に反映される場合もあるでしょうし（「前進」など)，複数の単語でもってロボットの動作一つを特定する場合もあるでしょう（「回れ　右」など)。障害物競走をするに際して，どのようなロボット動作が必要・可能なのか，また，それを操作するユーザにはどのような単語（コマンド）セットを提供するのが適当なのか。音声認識システムと，無線ロボット操作システムとの間のインタフェースを，各班でよく考えて製作して下さい。君らのセンスが問われます。

4　面白い話題提供　〜半日ほど，あれこれ思考してみましょう〜

今回の実験課題では「あ」という母音の音響的定義を行いました。「第一フォルマントが XX[kHz] 付近にあって，第二フォルマントが YY[kHz] 付近にあって … 」あるいは,「XX なスペクトル包絡特性を持つ音」，これが「あ」の音響的定義となる，と説明しました。もちろん,「あ」を作り出す時の口の形，長さは話者に依存するので，その音響的定義は，話者依存となってきます。

さて，父親と 3 歳の女の子とが会話しているシーンを思い浮かべて下さい。3 歳ともなれば色々父親にちょっかいを出して来ます。父親が言った言葉の中に知らない単語があり，それを女の子が真似て返しているシーンです。そう，新しい言葉を覚えているのですね。ここで「女の子は父親の生成した音声を真似ているのでしょうか？」と聞いたら，何と答えますか？「何言ってるの，真似てるに決まってるじゃない」と思った君，観察眼がまだまだ足りないです。女の子は決して父親の声をそのまま真似てなんかいません。父親の声を「物真似」「声帯模写」しようとなんかしていません。でも，父親の声の「何か」を真似ているはずですよね。一体全体，物理的（音響的）には，何を真似ているのでしょうか？

こう答えるかもしれません。「父親の声にあった知らない単語を平仮名に落として，その平仮名を一つ一つ自分の口で音に変えただけ，だから，父親の声の物真似なんかしなくても済む」と。でも，これは NG と

[16]技術的には当然，という意味です。機械による認識率の低い人の発話は，よく他人から誤解される，ということではありません。
[17]http://www.gavo.t.u-tokyo.ac.jp/~mine/jikken/speech_interface

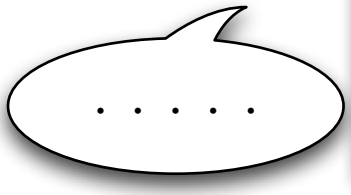
なります。彼らはまだ単語音声を平仮名に分けることができません。そもそも，日本語に 5 個の母音があって，それらが「あ」「い」「う」「え」「お」であることも知りません。「しりとり」だって出来ない状態にあります。こういう彼らの言語発達状態を考えると，「平仮名に落として，それを一個一個 … 」というのは，甚だ不適切な回答であると理解できるでしょう。で，最初の問いに戻ります。何を真似ているのでしょう？

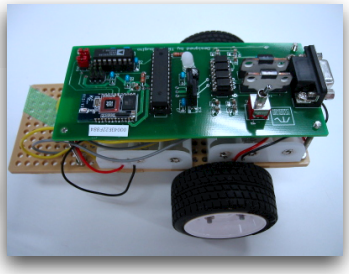
発達心理学（子供の認知能力の発達を議論する心理学）の教科書を紐解くと，例えば「子供は語全体の音形をまず獲得し，その後，個々の音（音韻／仮名）を獲得をする」と説明してある教科書もあります。となれば「語全体の音形」の物理的・音響的定義は何か？という問いになってきますね。

発達心理学者が言うように，3 歳の女の子は父親の発声の「語全体の音形」を真似ているとした場合，その「語全体の音形」には話者の情報は含まれていないことになります。でなければ声帯模写をし始めるはずです。結局「語全体の音形」は，話者の体格，年齢，性別には影響されない，ということになります。

発達心理学者にその物理的・音響的定義を聞いたことがあります。ですが，彼らの答えは「物理的定義など無い／分からない」です[18]。さてさて，前置きが長くなりました。面白い話題提供とは，本課題を通して身につけた知識を用いて，この「語全体の音形」の物理的定義を考えてみて下さい。話者の体格，年齢，性別に依存しない音形って一体何なのでしょう？一体全体，幼児は親の声の，何を，真似ているのでしょう？

ヒント：話者によって「あ」の物理的実体は異なります。「ある音」に対して，それが話者 A の声だとすると「い」と解釈され，話者 B の声だとすると「え」と解釈されることがあるのか？と言うと，答えは「あります。」日本語のように少ない母音数の言語では，(多数話者を考えた場合の）母音の音響的な重なりは大きくないですが，母音数が多い言語では重なりが一気に大きくなります。実は「音声の音響特性のみを使って，母音を 100%認識すること」は，(話者の違いを考慮すれば）原理的には不可能，というのが事実です。





[18]本当です。物理的定義を定めなくても，発達心理学としての議論ができるため，彼らは積極的に定義を求めようとはしません。

5　参考図書

- 「音の何でも小辞典」日本音響学会，講談社
- 「発声のメカニズム」萩野仁志・後野仁彦，音楽之友社
- 「ディジタル音声処理」古井貞熙，東海大学出版会
- 「音声情報処理の基礎」斉藤収三・中田和男，オーム社
- 「音声」中田和男，コロナ社
- 「音声・聴覚と神経回路網モデル」中川聖一・鹿野清宏・東倉洋一，オーム社
- 「音声情報処理」春日正男・船田哲男・林伸二・武田一哉，コロナ社
- 「音声認識システム」鹿野清宏・河原達也・山本幹雄・伊藤克亘・武田一哉，オーム社
- 「フリーソフトでつくる音声認識システム」荒木雅弘，森北出版
- 「99・9%は仮説」竹内薫，光文社
- 「科学者という仕事」酒井邦嘉，中央公論新社