

5 音声周波数の分析

5.1 発声の仕組み

図5-1は、人の発声に関わる器官を示したものです。人が発声する場合、肺から空気を送り出すことで、喉の奥にある声帯を振動させます。声帯は、2枚の薄い膜が横に並んだ構造になっており、呼吸をする場合などは、2枚の膜の間が開いています。この2枚の膜の間隔を狭め、その狭い隙間に肺からの空気を流すことで声帯を振動させ、音を発生させます。この声帯の2枚の膜の間の隙間のことを声門と呼びます。この時の声帯の振動数が基音の振動数に相当します。この時、同時に、基音の整数倍の周波数を持った倍音も発生します。音色の違いは、発声の際に、口の形や舌の位置等を調整することによって作り出しています。このように唇や歯などを使って様々な音を作り出すことを調音と呼びます。また、それらを唇や歯などの器官を調音器官と呼びます。口の中の空間である口腔の上側の部分は口蓋と呼ばれます。この部分も調音に利用されます。口の形や舌の位置等を変えることで、ある倍音は強調され、ある倍音は抑制されることで「あいうえお」といった音の変化が作り出されます。

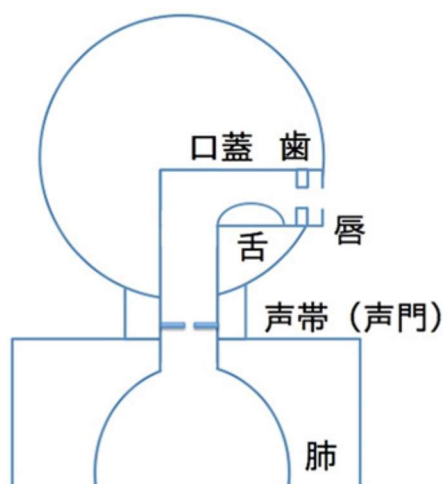


図5-1 発声器官

人が発する音声は、大きく分けて母音と子音の2つに分類することができます。母音は、声帯から出た音を口腔等で共鳴させることで発声します。声帯の振動を伴う音は、有声音と呼ばれ、母音は、有声音になります。これに対し、声帯の振動を伴わない音は無声音と呼ばれます。また、母音は、唇や歯で空気の流れを閉鎖したり狭めたりすること無く発声されます。唇や歯で空気の流れを閉鎖したり狭めたりすることで発声される音声は、子音と呼ばれます。母音は、ある程度の時間、発声し続けることができるという特徴があります。また、「あお」のように、他の母音に連続的に変化させることができるという特徴があります。

一方、空気の流れを妨げることで生成する音が子音です。子音には、さまざまなものがあり、いくつかの観点で分類することができます。表5-1は、日本語の主な子音の分類を示したものです。子音は、まず、調音位置で分類することができます。これは表の1行目に当たります。唇や歯などを使ってさまざまな音を作り出すことを調音と呼びますが、その調音する位置が調音位置です。調音位置としては、唇、歯、歯茎、口蓋、声門などがあります。口蓋は、口腔の上側の部分ですが、その前側の硬い硬口蓋と、後ろ側の柔らかい軟口蓋に分けることもできます。また、声帯の振動を伴う有声音であるか、声帯の振動を伴わない無声音かで分類することができます。これは、表の2行目に示しています。

また、音を発生させる方法により、破裂音や鼻音などに分類することができます。これは、表の1列目に示しています。例えば、破裂音としては、「b」や「p」や「d」や「t」があります。「b」や「p」は、一旦、口を閉じて、せき止めた空気を一度に吐き出すことで発声します。よって、調音位置は唇となります。また、「t」や「d」は、舌尖を上の前歯の付け根につけて急に離すことで発声します。よって、調音位置は歯、歯茎ということになります。また、発声する場合に声帯の振動を伴う音は有声音、声帯の振動を伴わない音は無声音と呼びますが、例えば、「b」や「d」は有声音、「p」や「t」は無声音になります。「g」や「k」は、舌を一旦、口蓋につけて、急に離すことで発音します。鼻音は、空気を鼻の奥の空間である鼻腔へ流し、共鳴させることで発声する音です。鼻音には、「m」や「n」といった音があります。日本語の「r」の音は、舌尖を歯茎につけて弾くことで発声する音で、弾き音と呼ばれます。また、摩擦音は、狭い隙間を作り、そこに空気を流すことにより発声する音です。例えば、「z」や「s」は、舌尖と上の歯茎の裏の隙間に空気を流すことで発声します。「z」は有

声音, 「s」は無声音です。調音位置が声門で無声音である「h」の音は, 「はひふへほ」の「は」の音です。この音は, 喉の辺りで空気を摩擦させることで発声しています。一方, 調音位置が唇で, 無声音の「Φ」で表しているのは, 「ふ」の音です。「ふ」の音は, 口先をすぼめて, そこに空気を流すことで発音します。英語の「f」は, 唇と歯の間に空気を流して発音する音で, 日本語の「ふ」の音とは, 違う音です。50音は, 子音が似ている音を縦に並べていますが, 「はひふへほ」の場合, 「は」と「ふ」で調音位置が異なっているわけです。「w」や「j」の音は半母音と呼ばれます。これは, 母音的な性質を持った子音です。母音へ移行する際の間過程で生成される音です。このような母音や子音を組み合わせることで, さまざまな音声が発声されます。

調音位置	唇		歯・歯茎		口蓋		声門
有声・無声	有声	無声	有声	無声	有声	無声	無声
破裂音	b	p	d	t	g	k	
鼻音	m		n		ŋ		
弾き音			r				
摩擦音		Φ	z	s	ʃ	ʒ	h
半母音	w				j		

表5-1 日本語の主な子音

5.2 周波数分析

「あ」という音声は, 大きな声で発声しても, 小さな声で発声しても「あ」と聞こえます。また, 高い声で発声しても, 低い声で発声しても「あ」と聞こえます。このことから, 音素の識別は, 音声信号から主に音色の違いを抽出することで行うことができると考えられます。

図5-2(a)は, 「あ」の波形で, 図5-2(b)は「う」の波形です。「あ」と「う」の音では, 音の波形が大きく異なっているのが分かります。このような波形の違いがあるのは, 基音と倍音によるものです。一般に, 最も振動数の低い音（基音）の整数倍の振動の音（倍音）は生じやすく, それらの音が基音に重なることで波形の違いが現れます。このため, このような波形の違いを分析することで音の識別を行うことができます。

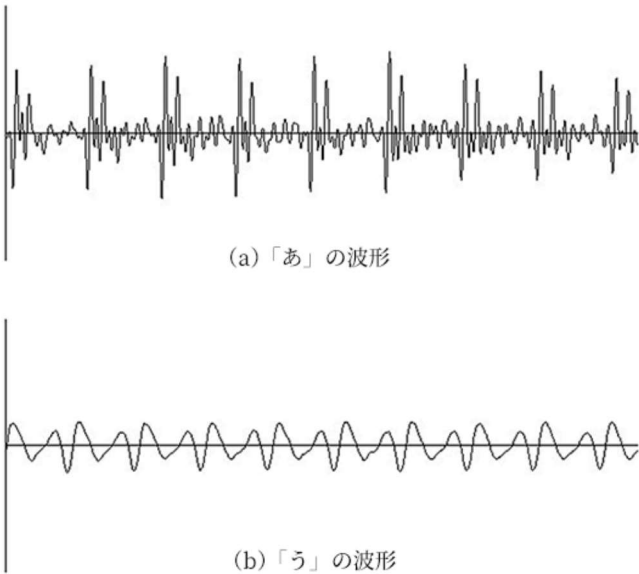


図5-2 波形の違い

ここで, 波の合成と分解について見てみましょう。図5-3では, 波の合成の例を示しています。図の左上には, 基音として $\sin(x)$ の波形が描かれています。また, その下には, 2 倍音として $\sin(2x)$, 同様に, 3 倍音として $\sin(3x)$, 4 倍音として $\sin(4x)$ の波形が描かれています。今, これらの波形を重み1, 1/2, 1/3, 1/4で足し合わせた場合を考えます。このような波形を作ると, 図の右側のような波形となります。多少でこぼこしていますが, 三角形の波が生成されていることが分かります。このような波形は, のこぎりの歯のような波形であることから, のこぎり波と呼ばれます。のこぎり波は, 弦楽器に音が近いことから, シンセサイザーで音を合成する場合のベースとしても利用されます。

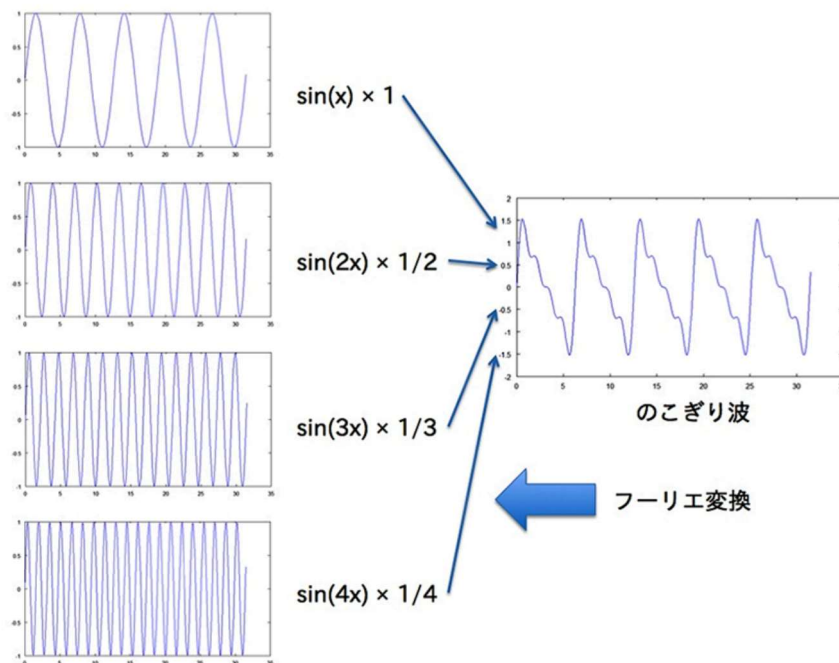


図5-3 波の合成と分解

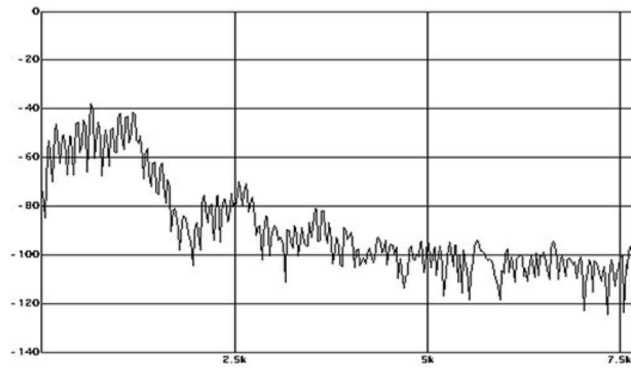
周波数分析では、逆に、与えられた信号にどのような周波数の波が含まれているかを分析します。すなわち、入力された信号が、複数の周波数が異なる波から合成されているとして、それらがどのぐらいの割合で含まれているかを分析します。周波数分析の手法には、さまざまな方法がありますが、広く利用されるものの1つに、フーリエ変換があります。

フーリエ変換は、与えられた信号を三角関数 ($\sin(kx), \cos(kx)$) の和として表現した場合の成分の強度を求めるものです。kの値が小さければ、振動数の低い低周波の振動に対応します。kの値が大きければ、振動数の高い高周波の振動に対応します。フーリエ変換した結果は複素数となりますが、その絶対値は、対応する周波数成分の強度となります。ただし、実際に、コンピュータ上でフーリエ変換の計算をする際には、通常、FFTと呼ばれる計算手法を利用します。FFTでは、横軸のデータ数が2のn乗であることを仮定します。この条件を満たせば、比較的簡単に高速にフーリエ変換を行うことができます。

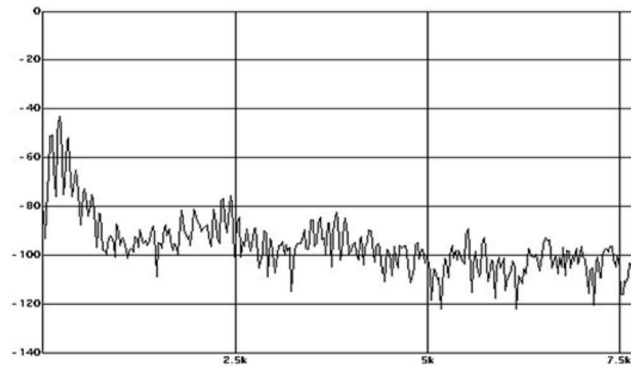
では、人の声をフーリエ変換してみましょう。例えば、「あ」の音をフーリエ変換して、それぞれの周波数の強度を求めると、図5-4(a)のようになります。横軸が周波数、縦軸がそれぞれの周波数の強度を表しています。この周波数スペクトルを見ると、周波数スペクトルには、大きく分けて2つの構造があることが分かります。大きなうねりの部分と、そこに重なる小刻みな振動の部分です。

小刻みな振動部分は、声帯の振動によって生成されます。声帯の振動数は、ピッチと呼ばれます。そして、この2倍音、3倍音というように、ピッチの整数倍の部分でスペクトルの強度が大きくなっているのが分かります。一方、大きなうねりの部分を見てみると、いくつかの周波数で山や谷になっているのが分かります。山となる部分は、ホルマントと呼ばれ、周波数が低い方から順に第1ホルマント、第2ホルマントと呼ばれます。また、谷となる部分は、アンチホルマントと呼ばれます。ホルマントやアンチホルマントは、声帯から発せられた音が、ある倍音は強調され、ある倍音は抑制されることで生じます。よって、これらの特徴を用いることで、発声された音素を推定することができます。

図5-4(b)は、「い」のフーリエ変換結果を示したものです。「あ」のフーリエ変換結果と「い」のフーリエ変換結果で、ホルマントやアンチホルマントの位置が大きく異なることが分かります。母音を認識する際には、例えば、第1ホルマントと第2ホルマントの位置が利用されます。



(a) 「あ」のフーリエ変換結果



(b) 「い」のフーリエ変換結果

図5-4 フーリエ変換結果

5.3 話者同定

「あいいうお」という音声は、発声する人が変わっても「あいいうお」と認識されます。これは、音声認識が、音声の中から、人によらない特徴を抽出することで行われるためです。一方、それ以外の部分には、個人に特有な特徴が含まれています。これらの特徴は、話者同定に利用できます。

話者同定に利用できる特徴としては、声の大きさ、話す速度、ピッチ、周波数スペクトルなどが考えられます。この中で、意図的に変えることができる声の大きさや話す速度は、話者同定には、必ずしも適していないと考えられます。このため、話者同定には、通常、ピッチや周波数スペクトルが用いられます。

ピッチは、音声波形をフーリエ変換して得られる周波数スペクトルの小刻みな振動部分の間隔に対応します。これは、声帯が振動する場合に、その基本振動数の2倍、3倍といった倍音も同時に発生し、その間隔が声帯の基本振動数と一致するためです。このピッチを抽出する方法として、周波数スペクトルを入力信号と考えて、フーリエ変換する方法があります。ただし、フーリエ変換とその逆変換は、ほぼ同様の性質を持っていることから、しばしば、フーリエ変換ではなく、フーリエ逆変換が用いられます。

このようにして求められた特徴は、スペクトラムの綴りを入れ替えてケプストラムと呼ばれます。また、ケプストラムの横軸は、周波数（フレクエンシー）の綴りを入れ替えて、ケフレンシーと呼ばれます。ケプストラムは、おおまかに言って、低ケフレンシー部分のいくつかの山と、高ケフレンシー部分の1つの山という構造を持ちます。ケプストラムの低ケフレンシー部分は、主に、周波数スペクトルの大まかなうねりの部分に対応します。ケプストラムの高ケフレンシー部分には、1つのピークが立ちますが、これが、ピッチに対応する部分になります。このピークは、周波数スペクトルの微細構造の間隔がピッチと等しくなることから生じるもので、このピークを抽出することで、話者のピッチを得ることができます。

一方、周波数スペクトルの大きなうねり部分は、口腔の形状などの影響を受けることから、他人が真似をしたり、自分で意図的に変えたりすることが難しい特徴となります。このため、周波数スペクトルは、話者同定に利用するのに適した特徴と言えます。周波数スペクトルは、もとの音声データから、数十msec程度の狭い範囲を選び出して作成されます。これは、分析する範囲内で、波形が大きく変わらないことが必要のためです。長い時間にわたる音声データの特徴分析に利用されるのが声紋です。声紋は、数十msec程度の狭い範囲ごとに周波数スペクトルを作成し、その強弱を明るさや色の変化として表現し、時間軸上に並べることで作成します。

図5-5は声紋の例です。横軸は時間で、左から右に、「あいいうお」と発声した場合に対応します。この声紋のパターンは個人ごとに異なり、意図的に変えることが困難であることから、話者同定に広く利用され、セキュリティ管理や犯罪捜査などでも利用されています。また、声紋は、「ビブラートがある声は心地よい」といった音楽の分析などでも利用されています。

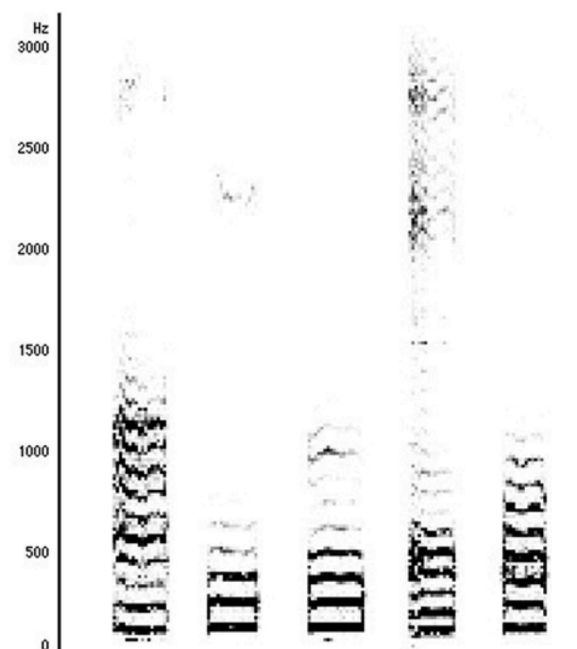


图5-5 声紋