作品名　iPhoneの音声認識の限界を探る

所属 愛媛県立八幡浜高等学校 学年　２　氏名　有本　陽菜，石岡　健大，平野　泰成

１．研究の要約

iPhoneは，音声を認識して文字に変換することができる。本論文では，iPhoneが認識することのできる最も単純な波形とその中に含まれる振動数を研究した。研究には，日本語の母音である５音を用い，男女５人ずつの波形データをフーリエ変換，逆フーリエ変換することによって最も単純な波形と振動数を明らかにした。母音の中では，「お」音が最も少ない振動数で表すことができ，２種類の振動数でiPhoneが正しく認識した。

２．研究の動機と目的

iPhoneの「Siri」やGoogleの「OK!Google」のように，スマートフォンやパソコンには音を文字として認識する音声認識システムが搭載されている。最近では，その精度はとても高く，文字を正確に認識することができる。このような音声認識は，音のどのような成分を読み取って文字に変換しているかを調べてみたいと考えた。

一般的に，音声認識には，音響モデルと言語モデルが使われている[1]。音響モデルは，音の波形をもとに音を特徴付ける要素（音素）に変換している。声の個性によっても正しく変換するために大量の音声データを使って統計的に処理を行っている。言語モデルは，単語を集めた辞書と単語の並び方を確率的に表現した辞書を用いて，音素の並びから文章を予測し変換している。

本研究では，音響モデルに着目し，iPhoneが認識する音声の限界の波形と振動数について明らかにした。

３．方法

男女５人ずつの声をUSB式マイクロフォンで録音し，波形を調べた。音響モデルの波形の変化での音声認識を調べるため，声は一文字ずつ録音し，波形をフーリエ解析することにより，音声に含まれる振動数を調べた。録音した音のデータ数は，CD音源と同じ1秒間に44100とし，0.5秒間の音をフーリエ解析した。最も簡単な波形を作るために，フーリエ解析によって求めた振動数をスペクトルが大きい順番に並べ直し，スペクトルが小さい振動数のスペクトルを0として，逆フーリエ変換を行うことによって，再び音を作成した。作成した音をiPhoneの音声認識を用いて文字にし，100%文字認識が成功する振動数だけを残した。

本研究では，日本語の母音である「あ」「い」「う」「え」「お」の５音について研究した。離散フーリエ変換と離散逆フーリエ変換は次式で与えられる。

$$F(k)=\sum_{n=0}^{N-1} f(n)e^{-i\frac{2\pi}{N}kn} \quad , \quad f(n)=\frac{1}{N}\sum_{k=0}^{N-1} F(k)e^{i\frac{2\pi}{N}kn}$$

この式を用いてフーリエ変換を行うプログラムを作成し解析に用いた。

４．結果と考察

1)最も少ないデータ数で認識することのできた母音の波形とパワースペクトル

i)「あ」音

図１に，10人の中で最も少ない振動数でiPhoneが正しく認識することのできた男性の「あ」音の波形を示す。縦軸が振幅で，横軸が時間を表している。横軸は0.02秒の間隔で表示している。図１の振動数とパワースペクトルを図２に示す。縦軸がスペクトルの大きさで，横軸が振動数を表している。縦軸と横軸はどちらも対数スケールで表している。スペクトルのピークは，振動数の小さい順に122Hz，244Hz，366Hz，486Hzのように，122Hzを基本振動とした倍音のところに存在している。 iPhoneが正しく認識するパワースペクトルの大きい振動数だけを残し，逆フーリエ変換を行うことによって作成した波形を図３に示す。縦軸が振幅で，横軸が時間を表している。横軸は0.02秒の間隔で，図１と同じ範囲の波形を表示している。波形は，図１と比較してもあまり変化していないように感じる。この波形を表現するために必要な振動数を表１に示す。データ数は13個であり，10人のデータ数の平均は23.6個で，母音の中で「お」音の次に少なかった。 図４は， iPhoneが認識する振動数だけを残

した「あ」音の振動数とパワースペクトルである。図４と表１から，iPhoneが正しく「あ」と認識する振動数の中には，122Hz，244Hz，366Hz，488Hz，608Hz，730Hz，974Hzの７つの振動数のピークが存在していることが分かる。これらの振動数は，基本振動122Hzの倍音であるが，854Hz付近の振動数が存在していないことから，7倍振動は「あ」音には関係ないことが分かった。残した振動数の中で，最もスペクトルの小さい振動数は974Hzであった。この振動数がなくなると正しく認識できなかったことから，974Hzが重要であることが分かった。

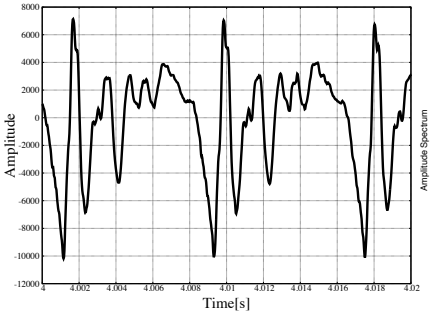


図１「あ」音の波形

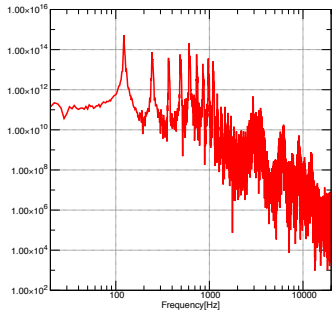


図２「あ」音に含まれる振動数とパワースペクトル

表1 「あ」音の中のスペクトルの大きい振動数

スペクトルの大きさ	振動数[Hz]
5.2596E+14	122
2.0736E+14	608
1.0697E+14	124
8.3601E+13	610
7.4273E+13	244
5.8174E+13	730
5.7121E+13	120
5.7118E+13	486
4.9864E+13	616
4.6489E+13	488
4.0946E+13	126
3.8700E+13	366
3.5975E+13	974

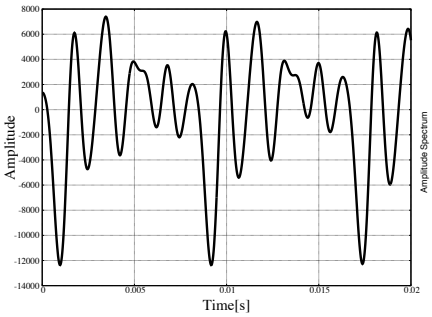


図３ iPhoneが認識する振動数だけを残した「あ」音の波形

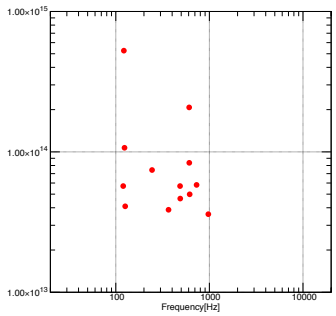


図４ iPhoneが認識する振動数だけを残した「あ」音の振動数とパワースペクトル

ii)「い」音

図5，6に，10人の中で最も少ない振動数でiPhoneが正しく認識することのできた男性の「い」音の波形と振動数とパワースペクトルを示す。図１の「あ」音の波形と比べると，振動数の高い波形になっているように感じる。図7，8にiPhoneが「い」音と認識することのできた波形と振動数とパワースペクトルの関係を示す。この波形を表現するために必要な振動数を表2に示す。データ数は11個であり，この人物は10人のデータのなかで一人だけとても少ないデータ数で認識することができた。10人のデータ数の平均は101個で，母音の中の真ん中で3番目に少なかった。図8のスペクトルのピークは，130Hz，260Hz，396Hzと130Hzを基本振動とする３種類の振動数で表すことができていた。

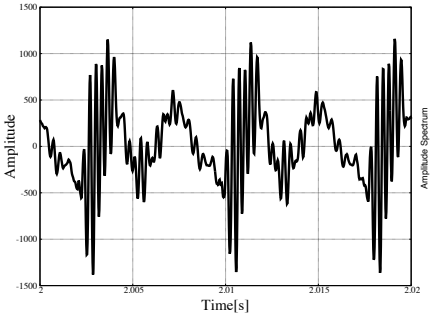


図5「い」音の波形

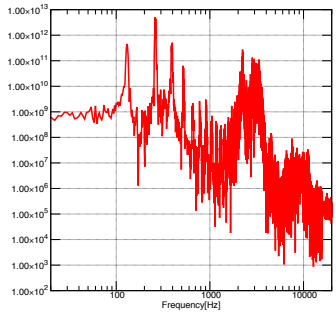


図6「い」音に含まれる振動数とパワースペクトル

表2 「い」音の中のスペクトルの大きい振動数

スペクトルの大きさ	振動数[Hz]
5.1531E+12	260
4.4583E+12	264
1.1672E+12	258
1.0736E+12	266
7.1498E+11	262
5.0705E+11	396
5.0241E+11	256
4.6623E+11	398
4.5767E+11	130
3.9226E+11	388
3.6602E+11	132

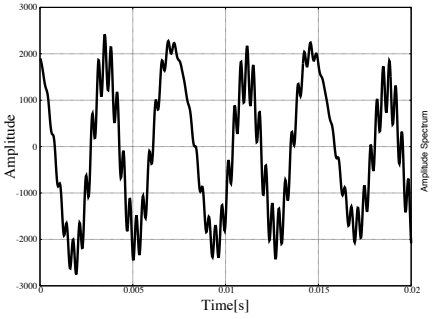


図7 iPhoneが認識する振動数だけを残した「い」音の波形

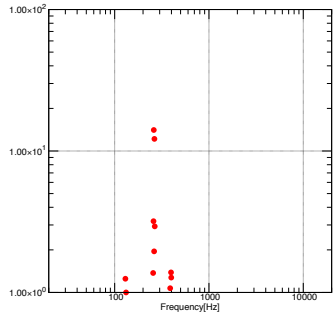


図8 iPhoneが認識する振動数だけを残した「い」音の振動数とパワースペクトル

iii)「う」音

図9，10に，10人の中で最も少ない振動数でiPhoneが正しく認識することのできた男性の「う」音の波形と振動数とパワースペクトルを示す。図１，5の「あ」「い」音の波形と比べると，単純な波形に感じる。図11，12にiPhoneが「う」音と認識することのできた波形と振動数とパワースペクトルの関係を示す。この波形を表現するために必要な振動数を表3に示す。データ数は52個であった。10人のデータ数の平均は282個で，母音の中で最も多い。図12のスペクトルのピークは，116Hz，234Hz，350Hz，468Hz，700Hz，934Hz，1402Hzであり，基本振動を116とすると2,3,4,6,8,12倍振動の値になっている。波形は単純に見えるが，構成する振動数は7種類であり，「あ」「い」音と比べると数が多い。また，データ数が52こと多くなっているのは，スペクトルのピーク付近の振動数も大きなスペクトルの値を取っているため，データ数が多くなったと考えられる。

表3 「う」音の中のスペクトルの大きい振動数

スペクトルの大きさ	振動数[Hz]	スペクトルの大きさ	振動数[Hz]
1.4876E+16	350	3.2184E+13	126
8.0192E+15	116	3.1036E+13	238
6.8056E+15	234	2.3963E+13	102
3.1677E+15	118	2.3865E+13	338
1.1313E+15	352	2.3439E+13	220
6.5505E+14	232	2.2081E+13	358
5.7462E+14	114	2.1287E+13	244
4.4474E+14	120	2.072E+13	128
4.3574E+14	354	1.9104E+13	246
3.4293E+14	346	1.691E+13	226
3.155E+14	348	1.6555E+13	100
2.5691E+14	236	1.6361E+13	130
2.5604E+14	112	1.6135E+13	98
1.0949E+14	110	1.5617E+13	242
1.0046E+14	122	1.4097E+13	248
8.7407E+13	108	1.3656E+13	96
8.2647E+13	230	1.2449E+13	132
7.6954E+13	228	1.2106E+13	934
7.2099E+13	344	1.1725E+13	94
6.7238E+13	124	1.1193E+13	468
5.4592E+13	342	1.087E+13	1402
4.8895E+13	240	1.0541E+13	134
4.8673E+13	106	1.0337E+13	92
4.4729E+13	104	1.0283E+13	362
3.6187E+13	700	1.0215E+13	702
3.2406E+13	224	9.7572E+12	250

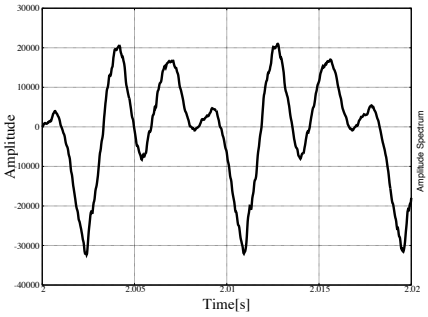


図9「う」音の波形

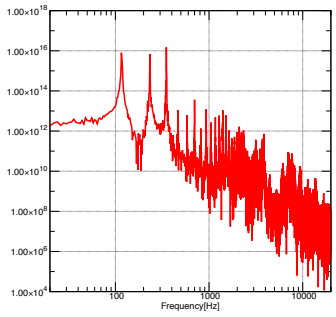


図10 「う」音に含まれる振動数とパワースペクトル

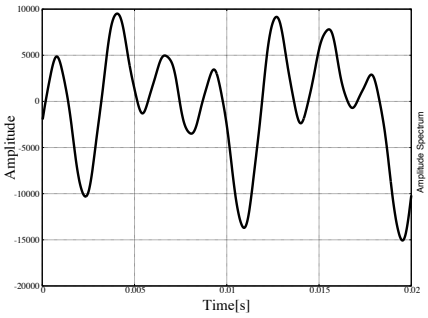


図11 iPhoneが認識する振動数だけを残した「う」音の波形

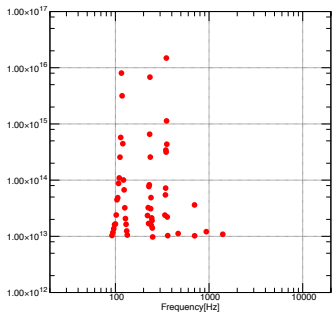


図12 iPhoneが認識する振動数だけを残した「う」音の振動数とパワースペクトル

iv)「え」音

図13，14に，10人の中で最も少ない振動数でiPhoneが正しく認識することのできた女性の「え」音の波形と振動数とパワースペクトルを示す。図15，16にiPhoneが「え」音と認識することのできた波形と振動数とパワースペクトルの関係を示す。この波形を表現するために必要な振動数を表4に示す。データ数は52個であった。10人のデータ数の平均は241個で，母音の中で2番目に多い結果であった。図16のスペクトルのピークは，238Hz，468Hz，716Hz，954Hz，2386Hzであり，480Hz，716Hz，954Hz，2386Hzは基本振動238Hzの2倍，3倍，4倍，10倍の振動数になっている。図16より，基本振動と２倍振動のスペクトルが大きくなっていることが分かる。

表4 「え」音の中のスペクトルの大きい振動数

スペクトルの大きさ	振動数[Hz]	スペクトルの大きさ	振動数[Hz]
4.1331E+14	480	8.4046E+12	712
2.1464E+14	242	8.2169E+12	470
2.0787E+14	476	8.0417E+12	246
2.0064E+14	238	6.9537E+12	722
1.937E+14	236	5.6333E+12	708
1.8195E+14	484	4.7681E+12	706
1.5282E+14	468	4.7639E+12	234
1.3968E+14	474	4.7108E+12	2386
1.1044E+14	240	4.5628E+12	726
9.1809E+13	482	4.5091E+12	248
8.067E+13	472	4.2288E+12	702
6.4217E+13	486	3.9181E+12	228
4.6946E+13	462	3.8449E+12	224
4.683E+13	466	3.3825E+12	698
3.2055E+13	478	3.3501E+12	728
3.0691E+13	230	2.7181E+12	458
2.7856E+13	460	2.5752E+12	2378
2.3372E+13	488	2.5275E+12	696
2.2474E+13	232	2.5229E+12	448
2.0522E+13	490	2.2307E+12	506
1.6811E+13	244	2.1605E+12	504
1.4727E+13	464	2.0841E+12	954
1.3103E+13	716	2.0799E+12	2392
1.0925E+13	456	2.0088E+12	704
1.0463E+13	454	1.9857E+12	450
1.0256E+13	718	1.9831E+12	692

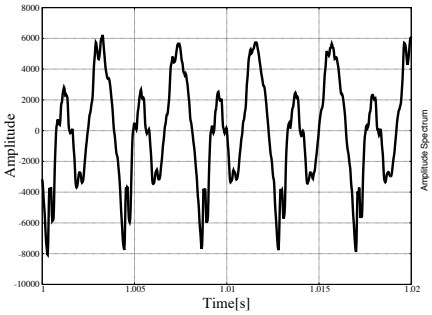


図13「え」音の波形

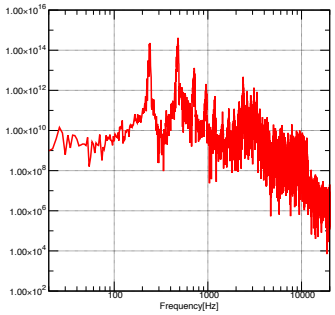


図14 「え」音に含まれる振動数とパワースペクトル

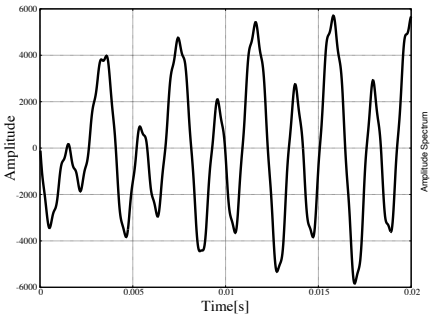


図15 iPhoneが認識する振動数だけを残した「え」音の波形

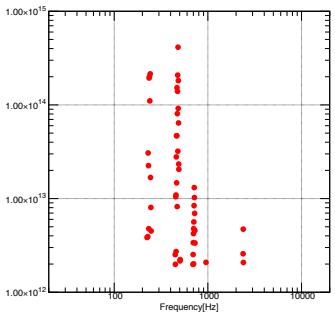


図16 iPhoneが認識する振動数だけを残した「え」音の振動数とパワースペクトル

v)「お」音

図17，18に，10人の中で最も少ない振動数でiPhoneが正しく認識することのできた女性の「お」音の波形と振動数とパワースペクトルを示す。「あ」～「え」音の波形と比べると，単純な波形に感じる。図19，20にiPhoneが「お」音と認識することのできた波形と振動数とパワースペクトルの関係を示す。この波形を表現するために必要な振動数を表5に示す。データ数は2個であった。2個のデータは，422Hzと634Hzであり，基本振動を211Hzとした場合の2倍振動と3倍振動であった。10人のデータ数の平均は22個で，母音の中で最も少ない結果であった。

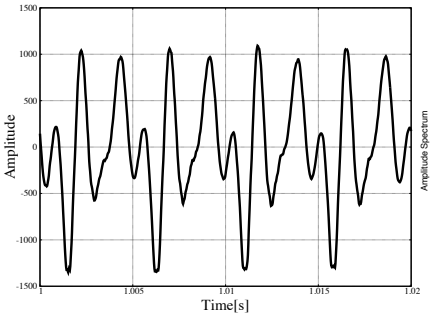


図17「お」音の波形

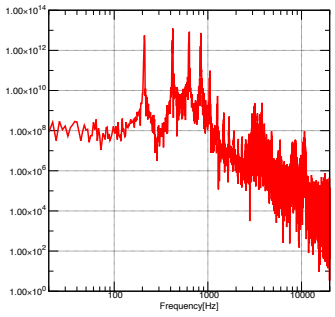


図18 「お」音に含まれる振動数とパワースペクトル

表5 「お」音の中のスペクトルの大きい振動数

スペクトルの大きさ	振動数[Hz]
1.2375E+13	422
8.4895E+12	634

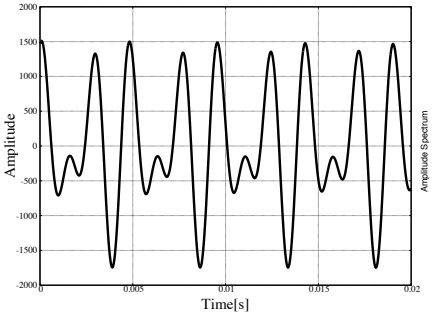


図19 iPhoneが認識する振動数だけを残した「お」音の波形

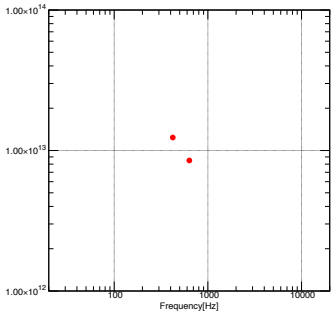


図20 iPhoneが認識する振動数だけを残した「お」音の振動数とパワースペクトル

２)10人の母音に含まれる振動数と基本振動との関係

i)「あ」音

図21に「あ」音の10人の基本振動に対する振動数の相対値を示す。縦軸は，iPhoneが正しく認識した振動数だけを残したスペクトルがピークの振動数を，基本振動数を１とし，相対値として表している。横軸は人物を表している。横軸は，1~5の赤色の点が女性のデータであり，6~10の青色の点が男性のデータを表している。10人全てのデータにおいて，基本振動の整数倍付近にデータが存在している。このことから，「あ」音を構成する振動数は基本振動の倍音で構成されていることが分かる。女性の場合，基本振動の値は190~240Hz程度，男性の場合は100~120Hz程度であった。「あ」音に含まれる倍音は基本振動の10倍までの間で構成されており，最も少ない人は，その中の4つの振動数だけで「あ」音を表すことができていた。図22に「あ」音の基本振動に対する振動数の相対値とスペクトルの相対値を示す。縦軸がスペクトルの大きさを表しており，対数スケールで表示している。最もスペクトルの小さい値を１とし，相対値で表している。横軸は，「あ」音の各人物の基本振動に対する振動数の相対値で，図21の縦軸と同じである。凡例の(1)~(5)の赤色が女性，(6)~(10)の青色が男性のデータを表している。スペクトルの大きい振動数には個人差があるが，女性の場合，３倍，４倍振動のスペクトルが大きくなっており，男性の場合は５倍，６倍振動のスペクトルが大きくなっている。女性の場合は，基本振動が200Hz程度なので，スペクトルの大きな振動数は，600~800Hz程度，男性の場合は基本振動が100Hz程度なので，500~600Hz程度の振動数が重要であると考えられる。(10)の男性が最も少ないデータ数でiPhoneが「あ」音を認識した。

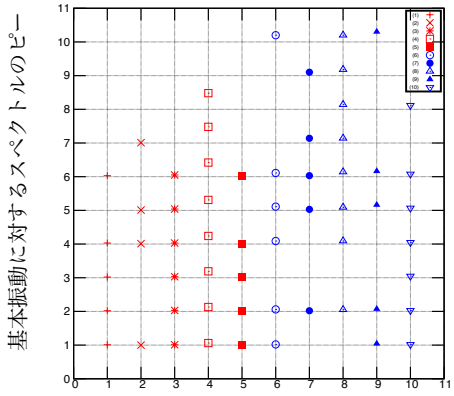
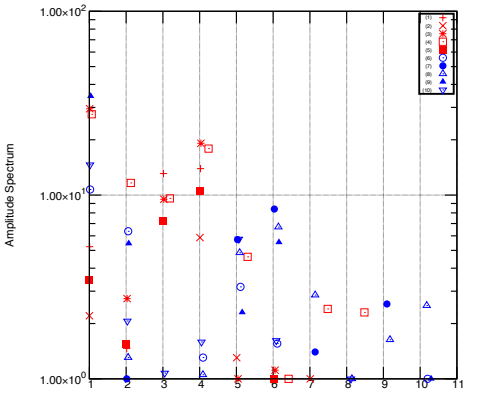


図21 「あ」音の各人物の基本振動に対する振動数の相対値



基本振動に対するスペクトルのピークの振動数

図22 「あ」音の基本振動に対する振動数の相対値とスペクトルの相対値

ii)「い」音

図23に「い」音の各人物の基本振動に対する振動数の相対値を示す。「い」音は「あ」音と比べて基本振動に対する振動数が大きい振動数まで存在している。「あ」音とは異なり，基本振動の整数倍以外の振動数も存在している。女性の場合，基本振動の5倍までの振動数の低い範囲と，10倍以上の振動数の高い範囲，男性の場合は，基本振動数の10倍までの範囲と15倍以上の範囲で別れて存在している。「い」音の基本振動の値は，女性の場合は200~280Hz，男性の場合は110~170Hz程度であった。基本振動数から，女性の基本振動の５倍までの範囲は，1000Hz程度であり，男性の基本振動の10倍までの範囲も1000Hz程度である。また，振動数の高い範囲は女性の場合は，基本振動の10~20倍，男性の場合は20~30倍の振動数であるが，基本振動から，女性は2000~4000Hz，男性は2000~3000Hzと同じ程度の範囲の振動数が存在している。図24に「い」音の基本振動に対する振動数の相対値とスペクトルの相対値を示す。基本振動の5倍までのスペクトルのピークが大きくなっている。女性の基本振動の10~20倍，男性の20~30倍の振動数のところのスペクトルも大きくなっている。(7)の男性が最も少ないデータ数でiPhoneが「い」音を認識した。

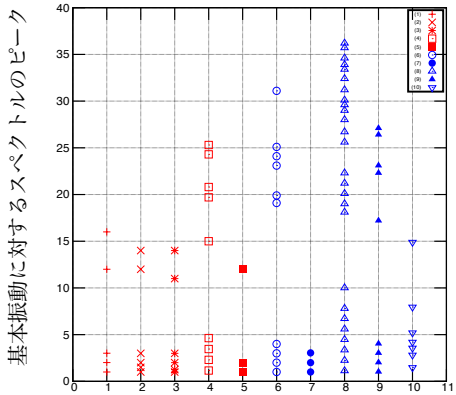
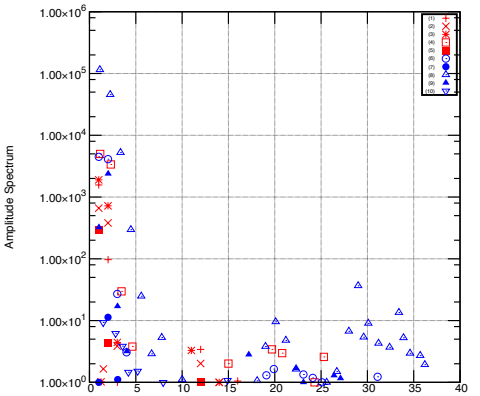


図23 「い」音の各人物の基本振動に対する振動数の相対値



基本振動に対するスペクトルのピークの振動数

図24 「い」音の基本振動に対する振動数の相対値とスペクトルの相対値

iii)「う」音

図25に「う」音の各人物の基本振動に対する振動数の相対値を示す。「あ」音とは異なり，基本振動の整数倍以外の振動数も存在している。「う」音の基本振動の値は，女性の場合は200~240Hz，男性の場合は100~120Hz程度であった。振動数の範囲は10人中8人は基本振動の20倍までの範囲で構成されているが，残り２人の最も大きい振動数は，女性は基本振動の40倍，男性は基本振動の70倍であったので，7000Hz~8000Hz程度と非常に大きい振動数まで含まれていた。図26に「う」音の基本振動に対する振動数の相対値とスペクトルの相対値を示す。振動数が大きくなるに従ってスペクトルの値も小さくなっている。(2)，(8)の人物のスペクトルは，基本振動の40倍まで大きなスペクトルが存在しているが，そのほかの人物は，基本振動の10倍までの間に大きなスペクトルが集中している。(7)の男性が最も少ないデータ数でiPhoneが「う」音を認識した。

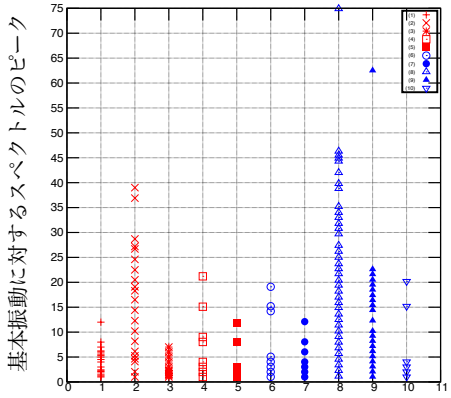


図25「う」音の各人物の基本振動に対する振動数の相対値

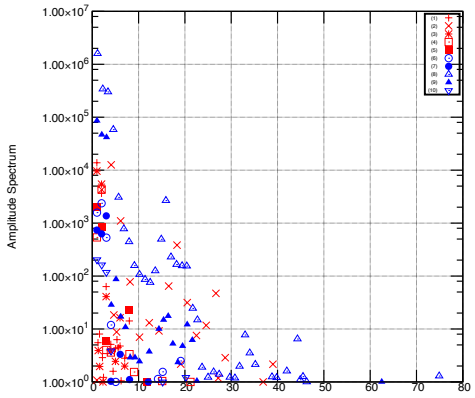


図26 「う」音の基本振動に対する振動数の相対値とスペクトルの相対値

iv)「え」音

図27に「え」音の各人物の基本振動に対する振動数の相対値を示す。「あ」音とは異なり，基本振動の整数倍以外の振動数も存在している。「え」音の基本振動の値は，女性の場合は200~280Hz，男性の場合は110~120Hz程度であった。女性も男性も基本振動の25倍までの範囲にバランスよく振動数が存在している。図28に「え」音の基本振動に対する振動数の相対値とスペクトルの相対値を示す。男性のスペクトルは，基本振動から10倍振動までの間が最も大きく，次に15~25倍振動のスペクトルが大きくなっている。(5)の女性が最も少ないデータ数でiPhoneが「え」音を認識した。

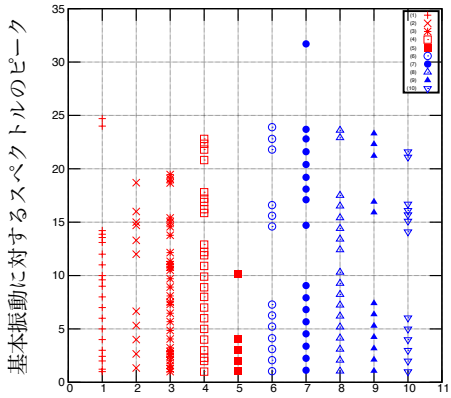


図27「え」音の各人物の基本振動に対する振動数の相対値

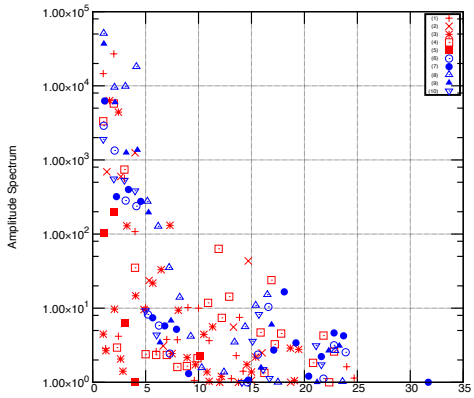


図28 「え」音の基本振動に対する振動数の相対値とスペクトルの相対値

v)「お」音

図29に「お」音の各人物の基本振動に対する振動数の相対値を示す。「あ」音と同様に，10人全てのデータにおいて，基本振動の整数倍付近にデータが存在している。このことから，「お」音を構成する振動数は基本振動の倍音が重要であることが分かる。「お」音の基本振動の値は，女性の場合は200~230Hz，男性の場合は100~110Hz程度であった。女性の場合

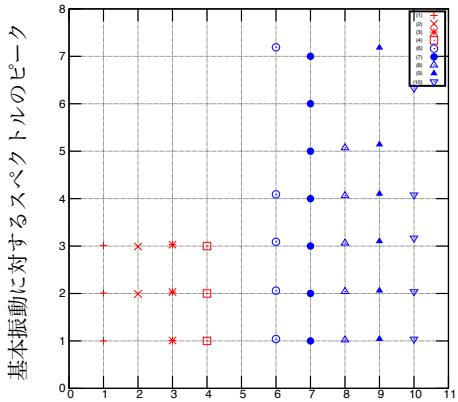


図29 「お」音の各人物の基本振動に対する振動数の相対値

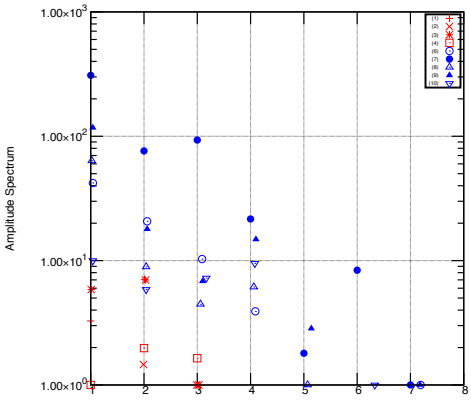


図30 「お」音の基本振動に対する振動数の相対値とスペクトルの相対値

は３倍振動，男性の場合は7倍振動までの振動数が存在している。基本振動の値から，女性も男性も600~700Hz程度までの振動数が重要であることが分かる。図30に「お」音の基本振動に対する振動数の相対値とスペクトルの相対値を示す。振動数が大きくなるに従って，スペクトルの大きさが小さくなっていることが分かる。(2)の女性が最も少ないデータ数でiPhoneが「お」音を認識した。

５．結論と今後の課題及び感想

1）結論

iPhoneの音声認識が認識できる限界の波形，振動数を求めるために，フーリエ変換を用いて解析を行なった。今回解析した男女10人の中で，最も少ないデータ数で認識することのできた人物の母音の中に含まれるスペクトルのピークは，全ての母音が基本振動の整数倍の値で表されていた。しかし，10人全てのデータを比較した時に，「い」「う」「え」音は，基本振動の整数倍以外の振動数もスペクトルのピークになっていた。女性の基本振動の値は200Hz程度で，男性の基本振動の値は100Hz程度であった。「あ」「お」音の振動数は，10人全てが基本振動の整数倍の振動数で表すことができ，基本振動の10倍までの大きさの振動数で表すことができる。基本振動の整数倍の組み合わせは，人物による個人差も含まれるが，女性の「お」音の場合は，基本振動の2倍である400Hz程度と3倍である600Hz程度の振動数だけでiPhoneが「お」音を正しく認識した。「い」「う」「え」音には，基本振動の20~40倍の大きさの振動数も含まれおり，低い振動数と高い振動数を同時に聴くことによって認識することができることが分かった。

2）今後の課題

本研究では，フーリエ変換して求めたスペクトルの大きい振動数だけを残すことによって単純な波形を作ろうと考えたが，フーリエ解析では，スペクトルのピーク付近の振動数も大きな数値を持っているので，iPhoneが認識できる範囲の中に含まれる振動数のデータ数も多くなった。今回の結果から，スペクトルのピークが重要であることがわかったので，スペクトルのピークだけを残して波形を作りiPhoneが認識することができるかを調べてみたい。また，「い」「う」「え」の３音には，基本振動の整数倍以外の振動数も含まれていると考えられるので，詳しく調べたい。

3）感想

今回の研究では0.5秒間のデータを解析した。0.5秒間のデータ数は22050個であるので，フーリエ変換のプログラムを1回回すのに22050^2通り計算する必要があった。そのため，1回プログラムを回すのに5分近くかかった。iPhoneが認識する限界の音を探すために，1音に対して何回もプログラムを回す必要があったので，とても時間がかかった。特に「い」「う」「え」音は，iPhoneが正しく認識するために必要な振動数の数が多く，限界を探るのに時間がかかり，苦労した。しかしその結果から，どの母音も基本振動の整数倍の振動数だけで認識されることがわかり，非常に興味深かった。

６．参考文献

[1] https://www.google.co.jp/amp/ascii.jp/elem/000/001/253/1253779/amp/