2. 音声圧縮の基本構造

2-1 聴覚の性質を利用した高能率圧縮のしくみ

正会員 宮坂 榮一†

キーワード：音符号号化，音声の高能率圧縮，低ビットレート，マスキング，ツビッカ

1. ま え が き

最近のデジタル音声放送やパッケージメディアにおける音声圧縮には、ほとんどが、非可逆な符号化方式が使われている。これらの圧縮符号化方式では、聴覚のマスキング特性をたくみに利用して、高能率化を実現している。音声圧縮符号化の基本的構成や主要な要素技術である「時間周波変換」、「再量子化」については2-2章で詳述されるので、ここでは、マスキングの性質について述べる。その理由は、こうした圧縮符号化方式においては、マスキング特性の設計上での扱いが、符号化音声の品質に大きな影響を与えるからである。

2. マスキング

2.1 臨界帯域幅

一般に、マスキングは、対象とする音の最小可聴値が他の音によって上昇する現象ないしはその上昇量(dB)と定義される。マスキングの量を強調するときは、「マスキング量」ということもある。マスクする音をマスキャーという。マスクされる音をマスキューということもあったが現在ではほとんど用いられず、単に「信号」ということが多い。

マスキングには、対象とする音が他の音に完全にマスクされて聞こえなくなる場合もあれば、聞こえにくくなる場合もある。後者の現象を“部分マスキング”という。

いま、周波数 f_A の純音Aと、中心周波数 f_B 、帯域幅 Δf の帯域雑音Bが同時に耳に提示されたとしよう。Bはマスキャーで、雑音の1Hz当たりのパワー N は一定とする。

まず、マスキャーBの帯域幅 Δf を固定し、信号Aのレベルを変えて、信号Aが丁度聞こえ始めるときのレベル(マスキングいき値)を求める。これをいろいろな帯域幅 Δf のマスキャーBについて行くと、図1に示す結果となる。

Δf がある値以上に大きくなると、信号Aのマスキングいき値はほぼ一定になってそれ以上は上昇しなくなる。

この現象は、聴覚系に“ある種”のバンドパスフィルタ作

図1 マスキャーの帯域幅とマスキングいき値の関係

用があることを仮定すると説明がつく。現在では、このフィルタのことを「聴覚フィルタ」と呼んでいる。図1で、それ以上広げても信号Aのいき値が上昇しなくなるときの Δf のことを、「臨界帯域幅」という。

2.2 聴覚フィルタの形

パターソンは、マスキャーとしてノッチ雑音(帯域阻止雑音)を用いて、純音のマスキングいき値を求めることにより、聴覚フィルタの形状を推定している。ただし、フィルタは周波数軸上対称であると仮定している。

ノッチ雑音の高域側および低域側成分は、周波数 f の純音をはさんで周波数軸上対称に配置される。雑音の中心周波数は f 、ノッチの幅を $2\Delta f$ とし、様々な Δf に対して、純音のマスキングいき値 P を求める。

Δf が小さいときは、聴覚フィルタを通過する雑音成分も多いので、マスキングいき値 P も高い。 Δf が大きくなるに従い P も減少していく。すなわち、マスキングいき値 P はフィルタ内の雑音成分のパワーに比例する。フィルタの形状は周波数 f の関数であるから $p(f)$ とすると、

$$P = k \cdot \sum p(f) \cdot \Delta f \quad (k: \text{係数})$$

これから、

$$p(f) = k' \cdot (\Delta P / \Delta f)$$

すなわち、 Δf に対するマスキングいき値の変化量 ΔP を求めれば、フィルタの形状 $p(f)$ が求められる。

2.3 2つの臨界帯域幅

バンドパスフィルタの帯域幅の定義として、等価矩形帯域幅 ERB をとったときの聴覚フィルタの帯域幅(単位Hz)を図2に示す。同図の点線は、ツビッカやシャープ等により求められた臨界帯域幅である。

†NHK放送技術研究所

"Basic Ideas of Low Bit Rate Audio Coding ; High Efficiency Audio Coding using Auditory Masking" by Eiichi Miyasaka (NHK Science and Technical Research Laboratories, Tokyo)

図2 2つの臨界帯域幅 (文献(1) 図3.8を改変)

図2をみると、1kHz以上の形状はよく似ているが、ERBの方が値が小さい。また、500Hz以下では相当の違いがみられる。その理由の1つは、ツビッカらは臨界帯域幅を求めたものの、フィルタの形状そのものは求めていることである。これが、フィルタの形状から求められるパターン等の臨界帯域幅と大きく異なる所以である。

2.4 マスキングパターン

前節で述べたように、マスクされる信号に關与する聴覚フィルタの形状は、マスクされる信号の周波数を固定し、マスキングの帯域や周波数を変えることにより推定した。

これとは逆に、マスキングを固定しておき、マスクされる信号Aの周波数を変えながら、信号Aのマスキングいき値を求めたものをマスキングパターンという。図3に、ツビッカ等により求められたマスキングパターンを示す。ただし、狭帯域雑音マスキングの中心周波数は1kHzである。

ところで、マスキング実験をしてみればすぐわかるが、マスキングパターンは、評定者によるばらつきが大きく、このようなきれいなパターンが得られることはむしろ少ない。評定者によってはマスキング量に10dB近くも差がでることがあり、内外の研究者もマスキング実験の困難さを述べているくらいである。しかし、ツビッカの優れているところは、こうした複雑怪奇なマスキングパターンを、大胆にも、きれいなパターンに整形したことにある。

もう一つの功績は、周波数軸の正規化である。今、図2に示した臨界帯域幅のフィルタが、聴覚系内の周波数軸に沿って隙間なく並んでいると仮定する。また、フィルタの最小中心周波数を50Hz、最大中心周波数を13.5kHzとし、フィルタの高域側の遮断周波数は、そのすぐ上のフィルタの低域側の遮断周波数と一致していると仮定する。この場合、0～15.5kHzの周波数帯域では、24個のフィルタが存在する。

ツビッカはこのフィルタの帯域幅である1臨界帯域幅を1Barkという単位で示した。この単位は、音の大きさのレ

図3 狭帯域雑音のマスキングパターン (文献(1) 図3.4を改変)

図4 図3をBarkで正規化した整形したマスキングパターン (文献(1) 図13.4を改変)

ベル(ラウドネスレベル)の単位にphonを取り入れた研究者、Barkhausenを記念して命名された。横軸の周波数fをBarkで正規化すると、図3は図4に、図5は図6になる。f→Barkへの変換は図2に基づいて行われる。

図6から、マスキングパターンはほとんど同じ形状になる。すなわち、横軸に沿って平行移動すれば、任意のBark値におけるマスキングパターンが得られる。結局、図4のパターンがあれば、任意の周波数の、任意のレベルのマスキングパターンが求められる。このモデルを用いることにより、高能率符号化モデルを大幅に単純化できるのである。

3. マスキングの高品質高能率符号化への応用

3.1 複合音のマスキングパターン

マスキングの性質を利用する高能率符号化では、マスキングは、符号化しようとする信号Sの周波数成分であり、マスクされる信号は、信号Sの周波数成分自身か、もしくは量子化雑音である。

まず、信号Sを周波数分析して得られる数多くの周波数成分一つ一つのマスキングパターンを求める。これにより、ある成分は、別の成分によって、完全あるいは部分マスクされる。部分マスクされる周波数成分に対しては、マスキングされる量に応じて、粗い量子化を行う。当然、量子化雑音が増えるが、周辺の成分によってマスクされる限り知覚されない。知覚されないぎりぎりまで量子化雑音を増やすことができるので、量子化ビット数を小さくできる。

図5 中心周波数の異なる狭帯域雑音のマスクングパターン
(文献(2)図3.3を改変)

図6 図5をBarkで正規化したもの(文献(2)図13.3を改変)

信号Sは、通常、多数の周波数成分を持つ複合音である。各成分のマスクングパターンを求め周波数軸に並べると、当然マスクングパターンは重なり合う。マスクングパターンが重なる部分はどのように処理すべきかが問題になる。線形加算できるのか？これに答えるには、マスクされるとはどのようなことなのかを、改めて検証しなければならない。

3.2 マスキングのメカニズム

聴覚神経系は周波数軸に沿って並んでいる多数の神経群で形成されている。それぞれの神経は、図7に示すように非常にシャープな周波数選択性を持つ。神経は、図中「応答野」と書かれた、極めて限られた範囲の信号周波数成分のみに反応し興奮する。最も感度がよい周波数のことを“特徴周波数”という。神経層はこの特徴周波数順に並んでいる。

興奮する神経の数(興奮量)を縦軸に、神経の特徴周波数を横軸にとって得られるパターンを“神経興奮パターン”あるいは単に“興奮パターン”という。マスクングパターンは、言い換えれば、マスキングが引き起こす神経系の興奮パターンを表していると考えることができる。

しかし、神経の応答野の両側には、抑制野という領域がある。今、2つの正弦波音A、Bが耳に入り、信号Aは応答野を、信号Bは抑制野を刺激したと考えよう。このとき、Bのレベルがある値以上になると、Aは完全に抑圧され、神経は興奮しない。これを聴覚神経層に拡大すれば、マスクングは神経系の抑圧によって起こる、ということになる。この抑圧過程は非線形である。すなわち、2つの音が同時に入ってきたときの出力は、各音が単独に入ってきたとき

図7 聴覚神経の応答野と抑制野(文献(1)図1.18を改変)

の出力の足し算にはならない。

ところが、マスクングでは、ほぼこの加算則が成り立つのである。ということは、この抑圧過程はマスクングにはほとんど寄与しないことを示唆している。

そこで聴覚心理的な仮説が考えられることになる。信号Aとして純音、マスキングBとして帯域雑音を考えよう。このとき、聴覚神経層のうち、信号Aの周波数に対応する場所を中心とする聴覚フィルタ内では、信号AおよびマスキングBが共に神経興奮を引き起こす。しかし、信号Aによる神経興奮の増加量が検知されない限り、その信号は聞こえないということになる。つまり、マスクングとは、マスキングBが信号Aを抑圧することによって起こるのではなく、(B+A)による神経興奮と、Bによる神経興奮との差が検知できないときに、マスキングBは信号Aをマスクしたというのである。これにより、各周波数成分のマスクングパターンは、単純に線形加算してよい、ということになる。

4. む す び

ここでは、時間軸上のマスクングについて述べる事ができなかった。これには、プレマスクング、ポストマスクングがあるが、特に、後者はマスクング量も大きく、影響する時間範囲も長いので重要な現象である。いずれにしろ、符号化音の品質に大きく影響するマスクングパターンの性質をよく知ることが、設計する上で極めて重要である。

(2001年8月23日受付)

〔文 献〕

- 1) B. C. J. ムーア著、大串健吾監訳；聴覚心理学概論、誠信書房(1994)
- 2) E. ツピッカー著、山田由紀子訳；心理音響学、西村書店(1992)
- 3) 宮坂栄一：“マスクングについて”、JASjournal, 7, pp.60-69(1997)

宮坂 栄一 1969年、横浜国立大学工学部電気工学科卒業。同年、NHKに入局。長野放送局を経て放送技術研究所に勤務し、聴覚、ハイビジョン音声方式、話速変換方式、ニュース音声認識の研究等に従事。現在、同所研究主幹、正会員、工学博士。