■2 群（画像・音・言語）－11 編（マルチメディア）

1 章 マルチメディアの基本概念

（執筆者：佐藤真一）［2011 年 2 月 受領］

■概要■

マルチメディアとは，旧来計算機で主として扱っていた数値情報やテキスト情報に加え，音響・音声情報，画像情報，映像（動画）情報などの，様々な種類の情報を統合して扱うことのできるメディアのことをいう．昨今では，テレビやゲームに加え，携帯電話や各種のインターネット上のサービスでも，数値やテキストに加え，音響，画像，映像情報が当たり前のように扱えるようになっており，これらはとりもなおさずマルチメディアということができる．

本編では，特に計算機システムによるマルチメディア情報の処理を念頭に置き，上記のようなシステムにおいてマルチメディアを扱うにあたって，必要となってくる様々な技術について解説する．本章では，続く章で解説される上記の技術を理解するに際し，必要となる様々なマルチメディアの基本概念の中でも，特に重要と思われる概念について概説する．

まずは，本編で扱うマルチメディアとはどのような概念であり，どのような特性があり，どのような技術的課題があるのかを明確にするために，マルチメディアそのものの定義について述べる．ついで，そもそも音響や画像・映像などのマルチメディア情報は，数値やテキストのようなそもそも人工的なメディアと異なり，元来実世界中で得られる情報であることを踏まえ，これを計算機システムで扱いうるデータとする際の基本原理などについて述べる．そして，計算機システムで扱いうるデータとなったマルチメディア情報を踏まえて，これらに対する最も基本的な演算として，類似度を考え，マルチメディアに対する類似度についての考え方や関連する話題について概説する．

【本章の構成】

本章では，上記のような話題について，1-1 節「マルチメディアとは」において本編におけるマルチメディアという概念を明確にし，1-2 節「マルチメディア情報の表現」において実世界情報としてのマルチメディア情報を計算機システムで扱うための基本原理について述べ，1-3 節「マルチメディア類似度」においてマルチメディア情報に対する最も基本的な演算であるマルチメディア類似度について概説する．

■2 群 － 11 編 － 1 章

1-1 マルチメディアとは

（執筆者：佐藤真一）［2011 年 2 月 受領］

まず，マルチメディアという言葉の一般的な定義について考える．マルチメディアは，旧来の計算機などで主として扱っていた数値情報やテキスト情報に加え，音響・音声情報，画像情報，映像（動画）情報などの，様々な種類の情報を統合して扱うことのできるメディアのことを指し，特にこれらのうちの一種類だけではなく，複数種類の情報を複合して扱う場合を特にマルチメディアという．更に，数値・テキスト・画像のような静的な情報のみの場合よりも，音響や映像のような動的で時間変化を含む情報を含む場合を指して特にマルチメディアと呼ぶことが多い．また，メディアという言葉が指すとおり，マルチメディアはある情報を受け渡す媒介であり，特に人間と計算機システムや通信システムなどのシステムとの間，あるいはそうしたシステムを介した人間と人間との間の情報を媒介する場合を主として指す．更に，人間を中心にして考えたときに，メディアの一種としてのマスメディアに見られるように，情報の流れが単方向に限られる場合があるが，特にマルチメディアの場合には情報の流れが双方向的であり，対話性が実現されていることが多い．こうした状況を踏まえ，本編では，システム，特に計算機システムによるマルチメディア情報の処理を念頭に置いて考えることとする．

次に，マルチメディアで扱う情報の種類について考える．上述のように，マルチメディアでは，数値やテキストに加え，音響，画像，映像などの複数の情報を扱う．初期の計算機では主として数値が扱われ，次いで文字・テキストが扱われるようになったが，実はこれらはいずれも極めて抽象度の高い情報である．そのため，計算機には扱いやすいが，人間にとってはその文脈がなければ解釈のできない情報である（180 と言われても，それが身長を表しているのか，体重を表しているのか，あるいは他の何らかの量なのかが分からなければ解釈ができない）．一方，音響・画像・映像などの情報は，実世界中にそのまま存在する情報を切り出したものであり，大変具体的な情報であり，ある程度の文脈情報も内在しており，実世界で生活しているわれわれ人間にとっては直感的で扱いやすい情報である．

しかしながら，これらの情報は計算機にとっては大変扱いにくい情報である．例えば，音響情報は各時点における音圧の時系列，画像情報は網膜などのスクリーン上のあらゆる場所の輝度（あるいは色）情報，映像情報は画像情報の時系列情報であり，これらを計算機で扱いうるディジタル情報にした場合，特に高品質のマルチメディア情報とすればするほど（時系列の標本化周波数を高める，スクリーン上の輝度値計測の解像度を上げるなど），そのデータ量は大量となる．そのため，数値やテキスト情報に比べ，マルチメディア情報を適切に扱うのは困難となる．しかしながら，昨今の計算機の計算性能の増大，メモリや外部記憶装置の容量の増大，ネットワークインフラストラクチャの整備による通信速度の増加などから，こうした問題は解決されつつある．それでも，マルチメディア情報の扱いにおいては，データ量を削減するための圧縮技術が必要不可欠となる．こうしてマルチメディア情報が適切にディジタル化されると，様々な種類が複合しているマルチメディア情報をディジタル情報として統合的に扱うことができる．加えて，蓄積・伝送が容易となる上，計算機による加工もできるようになり，様々なアプリケーションの可能性が広がることになる．マルチメディア

情報のディジタル表現については，1-2 節で扱う．マルチメディア情報の計算機による加工に関しては 2 章の一部で，マルチメディア情報の圧縮・蓄積・伝送に関しては 3 章で主として扱う．

さて，マルチメディア情報は人間にとっては直感で分かりやすい情報であると述べたが，これは人間がその認知能力を駆使し，マルチメディア情報を高度に抽象化して解釈することができるためと考えられる．そのため，人間が計算機システムを用いてマルチメディア情報を扱う場合にも，人間がマルチメディア情報を解釈しているのと同様に，抽象化したレベルの情報として扱いうることを期待している．しかしながら，ディジタル表現されたマルチメディア情報は極めて具体的なデータとなっており，それらに対応する，抽象化した，いわば意味レベルの情報とは大きな乖離があり，その対応関係は極めてあいまいであり，不明である．そのため，マルチメディア情報に対し，その抽象化した情報を付随情報としてあわせて取り扱うことが多い．こうした付随情報はメタデータと呼ばれる．今日広く利用されているマルチメディアを扱っている大部分のシステムは，メタデータを利用することによって初めて，マルチメディア情報の適切な利用を可能としている．例えば，ディジタルカメラで撮影した写真の管理システム,インターネット上の画像･映像の検索システムは端的な例であり，撮影時刻，GPS による撮影場所，人手で付けたタグなどの情報に強く頼っている．ゲームや放送映像制作支援システムなどでも同様である．一方，メタデータには，作成に手間がかかる，マルチメディア情報がもつ多義性・あいまい性に対応しきれない，メタデータ作成者の主観によるゆれが発生する，などの問題がある．そこで，計算機でマルチメディア情報を「解釈」して抽象化し，メタデータを自動生成する技術についても検討されている．こうした技術をマルチメディアの内容解析技術と呼ぶ．これは一般には極めて困難であるとして知られており，この困難さをマルチメディア内容解析におけるセマンティックギャップと呼ぶ．すなわち，マルチメディア情報として与えられている具体的なデータと，それらから人間が期待する高度に抽象化された情報との間に大きなギャップがあることがその困難さの主因である．この困難さを克服するために様々な研究が行われており，一部限定的ではあるが有効な技術も生まれてきているが，まだ完全な解決策は達成できておらず，いまだに大変活発に研究がなされている．マルチメディア情報の内容解析については 2 章で，セマンティックギャップについては 1-3 節並びに 2 章で，マルチメディア情報に対するメタデータについては 3 章で扱う．これらを踏まえた，実際の応用システム並びに関連する技術については，4 章で扱う．

■2 群 － 11 編 － 1 章

1-2 マルチメディア情報の表現

（執筆者：有木康雄）［2010 年 2 月 受領］

1-2-1 画像・映像の表現

(1) 光の物理量と心理物理量

光の量は物理的には，単位時間当たりのエネルギー（放射束 Q：単位〔W〕ワット）として測定されるが，人間にとっては目の特性により感じる明るさ（心理物理量）として観測される．人間の目が可視光の波長に対して感じる明るさは一様ではなく，波長 555 nm の黄緑色の光に対して感度が最大となる．この感度は，放射束 1 W（ワット）に対して単位時間当たりの光量（光束 F：単位〔lm〕ルーメン）で測定すると 683 lm として感じる．この波長 555 nm の色の光に対する最大の感度を 1 とし，それ以外の波長の感度を相対値で表したものを比視感度という．人間の目には個人差があるので，多くの人の感度特性の平均をとって国際照明委員会（CIE）で定めた比視感度を標準比視感度といい $V(\lambda)$ で表す [1]（**図 1・1**）.

物理量である放射束 Q の波長分布（分光放射束 $Q(\lambda)$：単位〔W/nm〕）が与えられると，心理物理量である光束 F は標準比視感度 $V(\lambda)$ を用いて，以下のように求められる [2].

$$F = 683\int_{v} Q(\lambda)V(\lambda)d\lambda \qquad (1・1)$$

光が対象物に入射すると，入射光束の一部は対象物の表面で反射されて反射光束となる．入射光束に対する反射光束の比を反射率 R，その波長分布 $R(\lambda)$ を分光反射率と呼ぶ．対象物からの反射光束 F_r は，入射光の分光放射束 $Q(\lambda)$ と対象物の分光反射率 $R(\lambda)$ を用いて，次式で与えられる [2].

$$F_r = 683\int_{v} Q(\lambda)R(\lambda)V(\lambda)d\lambda \qquad (1・2)$$

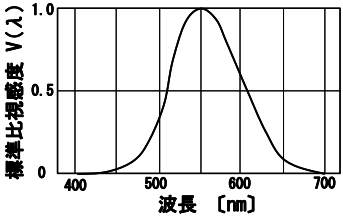

図 1・1 標準比視感度

(2) 色の物理量と心理物理量

任意の色光は三つの独立した色光の混合（加法混色）によって同じ色（等色）にできる．三つの独立した色光（原刺激）として，赤 R_0（700 nm），緑 G_0（546.1 nm），青 B_0（435.8 nm）の単色光を用い，任意の色光 C を等色したときのそれぞれの単色光の量（物理量の放射束，あるいは心理物理量の光束）を R，G，B とすると，次の等色方程式が成立する [2].

$$C = RR_0 + GG_0 + BB_0 \qquad (1・3)$$

可視波長域にわたって波長分布が一定である等エネルギースペクトル白色光（基礎刺激）に対して，加法混色により等色したときの原刺激の光束を l_R，l_G，l_B とする．その量は 1.0000 lm，4.5907 lm，0.0601 lm であり，明度係数と呼ばれている．任意の色光 C を等色したときの原刺激の光束が L_R, L_G, L_B であるとき，それらをそれぞれの明度係数で割った値 $L_R/l_R, L_G/l_G, L_B/l_B$ を，改めて R，G，B とおく．この値のことを三刺激値という．このとき，白色光は等色方程式において $R = G = B = 1$ で表現でき，任意の色光は R，G，B 3 次元空間のベクトルとして表現できる [3]．

種々の波長の単色光 1 W に対し，これと全く同じに見えるよう三刺激値の強さを調節して加える．こうして得られた三刺激値を，RGB 表色系のスペクトル三刺激値（等色関数）$\bar{r}(\lambda)$，$\bar{g}(\lambda)$，$\bar{b}(\lambda)$ と呼ぶ（**図 1･2**）．RGB 表色系は，スペクトル三刺激値に負量をもつため，負量をもたないよう線形変換した XYZ 表色系が，CIE によって設定されている．XYZ 表色系のスペクトル三刺激値 $\bar{x}(\lambda)$，$\bar{y}(\lambda)$，$\bar{z}(\lambda)$ のうち，$\bar{y}(\lambda)$ は標準比視感度 $V(\lambda)$ と一致するように作られている．

XYZ 表色系における三刺激値 X，Y，Z とスペクトル三刺激値 $\bar{x}(\lambda)$，$\bar{y}(\lambda)$，$\bar{z}(\lambda)$，分光放射束 $Q(\lambda)$ の関係は，次のように表される [1]．

$$X = k\int_v Q(\lambda)\bar{x}(\lambda)d\lambda, \quad Y = k\int_v Q(\lambda)\bar{y}(\lambda)d\lambda, \quad Z = k\int_v Q(\lambda)\bar{z}(\lambda)d\lambda, \qquad (1・4)$$

比例定数 k の値は，$\bar{y}(\lambda)$ が標準比視感度 $V(\lambda)$ と一致することと式(1･1)から，683 lm/W となる．

対象物からの反射光が目に入り色知覚を生じるのは，対象物によって分光反射率 $R(\lambda)$ が異なるからである．対象物からの反射光に対する三刺激値は，式(1･2)と同様にして以下のように求められる [4]．

$$X = k\int_v Q(\lambda)R(\lambda)\bar{x}(\lambda)d\lambda, \quad Y = k\int_v Q(\lambda)R(\lambda)\bar{y}(\lambda)d\lambda, \quad Z = k\int_v Q(\lambda)R(\lambda)\bar{z}(\lambda)d\lambda, \qquad (1・5)$$

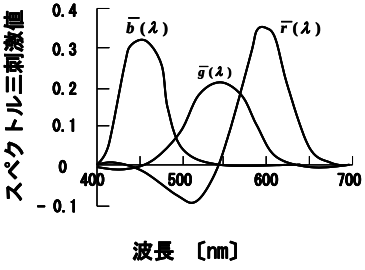

図 1･2　RGB 表色系のスペクトル三刺激値

(3)　電気量への変換

対象物や風景を画像として記録・蓄積するには，ディジタルカメラが使われる．ディジタルカメラは，レンズとシャッタ，固体撮像素子（エリアセンサ）と制御回路で構成されている．固体撮像素子は，撮影に必要な光電変換，電荷蓄積を行うフォトダイオードと，電荷転

送を行う電荷結合素子（Charge Coupled Device : CCD）を，1 枚のシリコン基板上に作った LSI である．ディジタルカメラでは，通常，数百万の画素を配列した固体撮像素子 1 枚を用いる単板方式と，3 枚用いる 3 板方式があるが，経済的な視点から単板方式が主流である．

対象物からの反射光が固体撮像素子に入射すると，フォトダイオードにおける光電効果により電子−正孔対が発生し，光の量に比例した電荷がフォトダイオードに蓄積される．その後，**図 1･3** に示すように，蓄積された電荷は一斉に垂直転送用 CCD に転送される．垂直転送用 CCD では，画素の電荷を一画素分上に送り，水平転送用 CCD に電荷をためる．水平転送用 CCD では，順次画素の電荷を左に送り増幅して出力する．水平転送用 CCD が空になると，垂直転送用 CCD で次の画素の電荷を 1 画素分上に送り，水平転送用 CCD に電荷をためる．この処理を繰り返す [2)]．

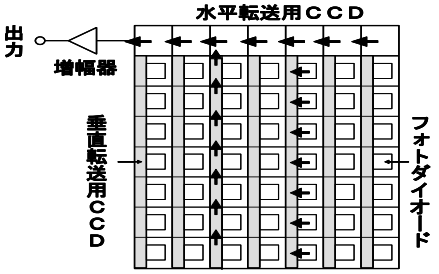

図 1・3　固体撮像素子の構成

カラー画像の場合には，フォトダイオードの感光部に色フィルタアレーを位置を合わせて配置する．フォトダイオードは，交互に異なる色信号を出力するので，どの画素においてもカラー出力できるように補間処理が施されている．

(4)　ディジタルへの変換

写真のようなアナログ画像をコンピュータで処理するためには，ディジタル画像に変換する必要がある．このアナログ画像からディジタル画像への変換では，「標本化」と「量子化」という二つの操作が行われる（**図 1･4**）．標本化は，空間的に連続した画像を離散的な画素の集合に分割する処理である．細かく標本化するほど，画素の大きさは小さくなり，単位面積当たりの画素数が増えるため，解像度が高くなって精細な画像を表現することができる．縦方向に N 画素，横方向に M 画素標本化した場合には，ディジタル画像は $N \times M$ 画素の 2 次元配列となる．

ディジタル画像への変換では，元のアナログ画像に含まれている特徴や情報が，ディジタル画像に正しく保存されている必要がある．すなわち，ディジタル画像から元のアナログ画像を正確に再現できることが必要である．これを保障する理論が標本化定理であり，「アナログ画像が含んでいる最高周波数の 2 倍より高い周波数でアナログ画像を標本化すると，ディジタル画像から元のアナログ画像を完全に復元できる」ことを保障している．この標本化定理を無視して標本化すると，得られたディジタル画像には周波数成分の重なり，あるいは折り返しが入り込んで，元のアナログ画像が正しく再現されなくなる．この現象をエイリアシ

ングという．

量子化は，その画素の濃淡を離散的な整数値に変換する操作である．量子化の際に割り当てられた濃淡の量子化の数を諧調数と呼ぶ．量子化を細かくして諧調数を増やすほど階調表現が豊かになり，正確に元のアナログ画像を再現できる．通常，1 画素当たり 0～255 の濃淡値をもつ 256 階調（8 ビット）が用いられる．ディジタル画像は，アナログ画像の近似値であるので，量子化により誤差を生じる．この誤差は，量子化誤差と呼ばれている．カラー画像の場合には，赤（R），緑（G），青（B）ごとに標本化，量子化が行われるため，1667 万色（$2^8 \times 2^8 \times 2^8$）を表現することができる．

ディジタルカメラでは,通常,数百万の画素を配列した固体撮像素子が組み込まれており，標本化と量子化が実装されている．

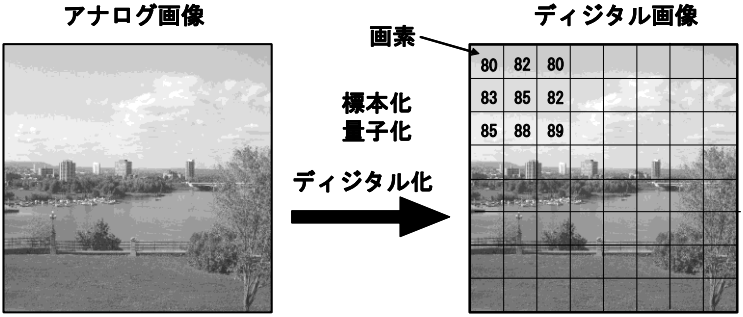

図 1・4　画像のディジタル化

1-2-2 音響信号の表現

(1)　光の物理量と心理物理量

弦や膜が振動すると空気も振動し，密な部分と疎な部分が作られ，音波となって伝わっていく．このように，振動方向と波の進む方向が同じである波を縦波（疎密波）という．これに対して，海面に石を投げ入れると，上下に振動しながら波となって伝わっていく．このような，振動方向と波の進む方向が直交する波を横波という．

海面のような横波の場合，時間 t を固定して見ると，波 $u(x)$ は海面における波の模様を表している．波の山から山までの距離を波長λ（単位〔m〕メートル）と呼ぶ．逆に位置 x を固定してみると，波 $u(t)$ はその位置における波の時間的変位（上下）を表している．波の山から山までの時間を周期 T（単位〔s〕秒）と呼ぶ．

音波のような縦波の場合も，時間 t を固定して見ると，**図 1･5**(a)に示すように，波 $u(x)$ は空間における空気の疎密の模様を表している．空気密度の高いところを縦軸のプラスに，空気密度の疎なところを縦軸のマイナスにとると，縦波 $u(x)$ も図 1･5(b)のような波形として表せる．逆に，位置 x を固定してみると，波 $u(t)$ は，その位置 x における空気の時間的変位（左右）を表している．空気の進行方向への変位を縦軸のプラスに，逆方向への変位を縦軸のマイナスにとると，縦波 $u(t)$ も図 1･5(c)のような波形として表せる [5]．したがって，縦波も横波も，時間 t と位置 x の関数として，$u(x, t)$ として表現できる．

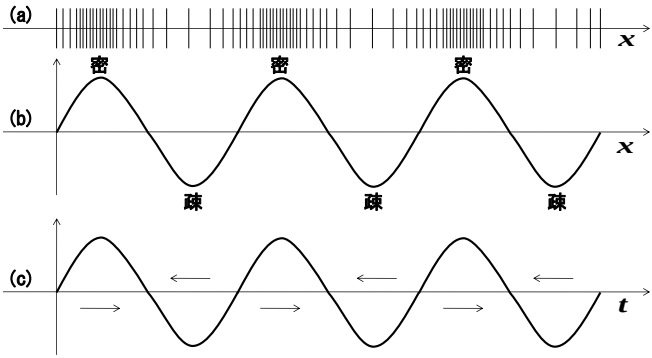

図 1・5　音波の波形表現

1 秒間に波が振動する回数を周波数 f（単位〔Hz〕ヘルツ）と呼ぶ．周波数 f と波長 λ は反比例するが，波の伝わる速さを速度 v（単位〔m/s〕）とすると，$v = f\lambda$ の関係がある．周波数の逆数，すなわち 1 回の振動に要する時間は周期 $T = 1/f$ である．振動の中心から山の高さ，谷の深さを振幅 A と呼ぶ．

空気中の波の速度（音速）は，0 ℃のとき，331.5 m/s であり，気温の上昇とともに早くなっていく．t ℃のときの音速 v は，$v = 331.5 + 0.6\,t$ であり，15 ℃のときの音速は約 340 m/s になる．このとき，音の波長 λ は，$\lambda = v/f \approx 340/f$〔m〕であり，1000 Hz の音でほぼ 34 cm である．人が聞き取れる可聴音 20 Hz～20,000 Hz では，ほぼ 17 cm～17 m になる．

(2)　音の心理物理量

人が音を聞いて感じる感覚は，音の「大きさ」，「高さ」，「音色」の三つである．「高い音」とは，周波数 f が大きい音であり，「大きい音」とは振幅が大きい音である．音の大きさは，一般に振幅の 2 乗と周波数の 2 乗に比例する．したがって，同じ高さの音であれば，振幅の大小が音の大きさを決めることになる．同じ高さの音でも，楽器によって「音色」が異なる．同じ高さなので波長は同じであるが，波形の違いが音色の違いとして知覚される．すなわち，周波数成分の相違により音色の違いが生じる．

人の耳が感じる「音の大きさ」は，同じ周波数であれば，音圧 p（単位〔Pa〕パスカル）が大きいほど，大きな音として認識される．人が認識できる音の音圧は，範囲があまりに大きいため，基準音圧 $p_0 = 20\times10^{-6}$〔Pa〕との比の常用対数により，音圧レベル $Lp = 20\times\log_{10}(p/p_0)$（単位〔dB〕デシベル）として表現される．

一方，人が聞くことのできる最小の音（最小可聴音）の音圧レベルは，**図 1･6** に示すように周波数によって異なっている．低い周波数では，音圧レベルが大きくないと聞こえない．そこで，音の周波数を変化させたときに，聴覚的に聞こえる音の大きさ（ラウドネス）が等しくなる音圧レベルを，図 1･6 のように描くことができる．これが等ラウドネス曲線である．等ラウドネス曲線では，1000 Hz における Lp〔dB〕の音圧レベルの音を Lp ホンとしている．最小可聴音は，4 KHz 付近で最小の音圧レベルとなっており，このことは，人間の耳は 4 KHz 付近で最も敏感であることを表している [6]．

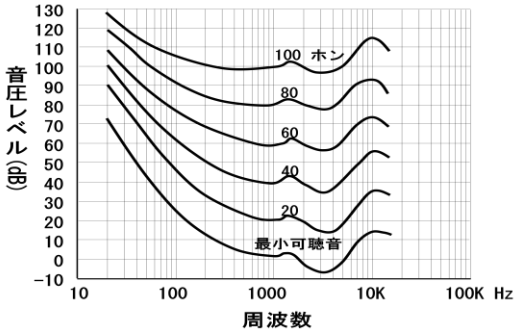

図 1・6　等ラウンドネス曲線

(3)　電気量への変換

音波を電気信号に変えるには，構造が単純で，平坦な周波数特性が容易に得られるコンデンサマイクロホンがよく用いられる．コンデンサマイクロホンの構造を，**図 1・7** に示す．位置 x における音波 $u(t)$ は，その位置 x における空気の左右方向の時間的変位を表しており，この空気の変位が，コンデンサマイクロホンの振動膜を振動させる．この結果，振動膜と電極で構成されるコンデンサの容量が変化する．

振動膜と電極との距離を d，振動膜の変位が電極に近づく方向を X の正の方向とし，振動膜の微小変位を ΔX とすると，コンデンサの容量変化に伴う開放起電力 ΔE は，$\Delta E = -E_0 \Delta X/d$ として計算できる．E_0 はコンデンサマイクロホンにかけられている電圧である．これより，振動膜の微小変位 ΔX，すなわち音圧に比例した出力電圧を，音響信号として取り出すことができる [7]．

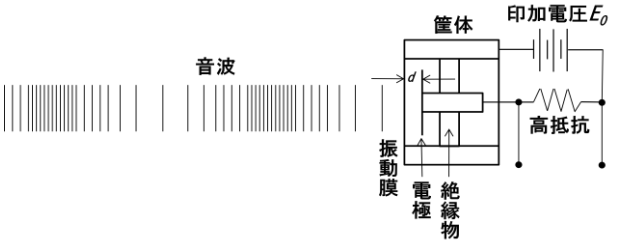

図 1・7　コンデンサマイクロホンの構造

(4)　ディジタルへの変換

コンデンサマイクロホンで収録したアナログの音響信号は，コンピュータで処理するために，ディジタル信号に変換する必要がある．ディジタル信号への変換では，画像のディジタル化と同じく，時間的に連続した音響信号を離散的な標本点の系列に分割する標本化と，音響信号の振幅値を離散的な整数値に変換する量子化が行われる．

■参考文献

1) 金出武雄, 坂井利之, 山口昌一郎, 栗田正一, 高梨裕文, “光と画像の基礎工学,” 電気学会大学講座, 電気学会, pp.193-204, 1984.
2) 長谷川伸, “画像工学,” (電子情報通信学会編), pp.13-97, コロナ社, 1983.
3) 南　敏, 中村　納, “画像工学―画像のエレクトロニクス―,” (テレビジョン学会編), コロナ社, pp.5-23, 1989.
4) 太田　登, “色彩工学,” 東京電機大学出版, pp.63-77, 1993.
5) 都筑卓司, “なっとくする 音・光・電波,” 講談社, pp.9-29, 1998.
6) 吉澤純夫, “音波シミュレーション入門,” CQ 出版社, pp.15-25, 2002.
7) 小林健二, “音・振動による診断工学,” (日本音響学会編音響テクノロジーシリーズ), コロナ社, pp.84-86, 2000.

■2 群 － 11 編 － 1 章

1-3 マルチメディア類似度

執筆中