量子化と包摂（ユニフィケーション）

東京工業大学情報理工学研究科
太田昌孝

概要

最初の JIS 漢字コードである JIS C 6226-1978 では一つのコードに複数の字体が対応し、包摂と呼ばれるが、包摂の工学的に適切な扱いのためには、包摂を入力における量子化誤差と出力における偏り（誤差）に分ける必要がある。文字コードと文字の入出力を電圧の AD/DA 変換と対比した結果、現行の JIS 漢字コードである JIS X 0208:1997 には、入力における偏り（誤差）を考慮していない、出力の許容誤差が不必要に厳しい、などの各種の問題があることがわかった。

Quantization and Unification

Masataka Ohta
Graduate School of Information Science and Engineering
Tokyo Institute of Technology

Abstract

JIS C 6226-1978, the first JIS Kanji code, maps multiple glyphs to a single code, which is called unification, proper engineering treatment of which requires separation of unification into quantization error on input and offset (error) on output. By comparing character input and output to/from character code with AD/DA conversion of voltage, it is found that the current JIS Kanji code JIS X 0208:1997 have various problems such as ignorance on input offset (error) and unnecessarily strict error allowance on output.

1.はじめに

文字コードとは、ざっくばらんには、図形である文字をコード化する仕組みと言えるが、文字コードには文字の持つ文化的側面があり、これを割り切って工学的に扱えるような概念に落とし込むことは必ずしも容易ではない。例えば、文字コードとはどうあるべきで、平テキストと構造化テキストの違いは何かという問題がある。これについては、筆者は[1]で、「文字の入出力はアナログ的な図形として行うことも容易だが、検索ではコード化されていることがパターン認識問題を避けるために本質的である」という考察の元、実用的な検索のためには文字コードは有限状態でなければならないことと、平テキストと構造化テキストの差も状態が有限かどうかにあると論じ、一応の解答を与えたつもりである。

文字コードに関して残る大きな問題は、JIS 漢字コード（以後 JIS という）の制定の際に導入された「包摂」という概念の工学的扱いである。制定当時の JIS[2]の解説ではまだ包摂という言葉は使われていないが、「漢字の異体字の取扱い」として「一つの符号位置に表示されている一つの字形は、ある範囲の変異（ゆれ）を許容し、それらを代表する一例であると考えるべきである」とあり、パターン認識に関係するようだ。ただ、このような文学的表現は、工学的な議論に耐えるような明確な定義ではない。実はこれだけならあまり問題はなく、「ちゃんと定義できてないね」と笑って済ませることができたが、Unicode[3]では日中韓の対応はするが字体の異なる漢字に同じコードを割り当てた上、その正当性の根拠を JIS の包摂だとしたため、日中韓の漢字が混在する環境では工学的にわけのわか

らない文字コードとなっている。

JIS も Unicode の影響を受け、[4]では、包摂を「複数の字体を区別せずに、それらに同一の句点位置を与えること」と定義しているが、その工学的意味は相変わらずよくわからない。ところが、基本部分がこのようにあいまいなままで「各句点位置では、そこに包摂される字体は相互に区別されない」とし、さらに、各コードが包摂する字体の範囲を詳細に規定したため、やはり工学的にわけのわからない状態である。

[2]にはない例だが人名に関わるため包摂で最も有名な例として「髙（くち）」と「髙（はしご）」があり、以後もこの例を多用する。[4]の包摂基準では、これらの文字は同じ符号位置とされる。そこで、「髙（はしご）」はその符号位置の文字として入力されるべきであり、その符号位置を「髙（はしご）」と出力してもよいこととなっている。しかし実際には「髙（はしご）」は外字として別に扱われることが多く、また当該符号位置を「髙（はしご）」と出力するような装置は「髙橋」という名前の人が特別に改造した装置でもない限り、ありえない。[2]の規定は現実と乖離しているのである。それにもかかわらず[5]で第三水準、第四水準の漢字を規定した時、[2]の包摂基準では既に「髙（はしご）」は[2]に含まれているという理由で、「髙（はしご）」が新たな漢字として導入されることはなかった。

そこで本稿では、包摂について工学的な割り切りをする。具体的には、包摂を図形を文字とする際のデジタル化に伴う量子化誤差と、文字を出力する際の出力偏りと捉えることで、包摂が工学的に扱えること示し、文字コードのあるべき姿について電圧の AD/DA 変換と対比しながら示す。

なお、本稿では、文字といえばいわゆる「図形文字」のことであり、いわゆる「制御文字」については論じない。

2.アナログとデジタル、図形と文字

アナログとデジタルの違いは、アナログでは細部の差に意味を求めるが、デジタルではある程度で切り捨てることにある。信号に周波数帯等の特性が信号と同じ雑音が乗った時、アナログではどうしようもないが、デジタルでは多少のノイズは切り捨てにより除去できる。そこで、情報のデジタル化により、長期間の保存や多段の処理の繰り返しによる情報の劣化を防ぐことができる。

なお、デジタル化された有界な連続量は有限個のビットでコード化することができるが、コード化されていないとデジタルではないというわけではない。例えば、書道の文字はアナログであるとしても、言語情報を表現するための文字は、もともとデジタルである。音素もデジタルであるし、そもそも言語自体がデジタルである。言語や音素や文字は、声の高低や強弱、書体や筆法の差、筆のかすれなどの細かな差異を無視して、多少の声嗄れ、紙の虫食い、石碑の磨耗等の雑音にも耐え、言語や音素や文字として情報が伝達できる。

文字はデジタルであるため、前後関係により図形としての形が変化しない場合、文字の数だけの図形で表現できる。変化が限定的である場合も、同様である。この結果生まれたのが活字であり、タイプライターである。

このように、文字や活字は本質的にデジタルであり、文字コードとは、文字に番号を振ったもの（正確には文字の並びを番号の並びに変換する規則だと怒る人もいるが、エンコード効率を無視して有限状態を展開してしまえば文字に番号を振るのと同じことなので、あまり気にする必要はない）にすぎない。すると、文字や活字やタイプライターの長い歴史の後に文字コードが生まれ、その文字コードが JIS において漢字化された瞬間、唐突に包摂という概念が出現するのはおかしい。包摂という概念が適切なものならばその概念は文字全般に適用可能なものであるはずだし、不適切なものだとしてもその元となる概念は文字全般に内在するはずである。

包摂について、漢字の字種の多さに由来してコード化に伴い発生する現象ではないかという誤解もあるが、漢字やその活字がもともとデジタルである以上、いまさらのコード化は原因ではない。実際、ラテン文字においても、JIS の包摂と同様の現象は当たり前に起きている。例えば、タイプライターの小文字の「l」と数字の「1」や大文字の「O」と数字の「0」は、しばし

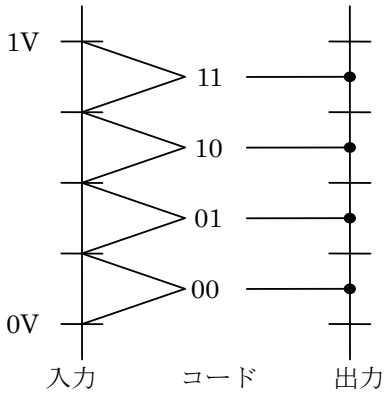

図 1 理想的な AD/DA 変換

ば同じキーが使われるが、これは[4]の定義によれば立派な包摂である。別の例として、文字コードが 6 ビットだった時代にはラテン文字は大文字だけだったが、小文字を含む文書も、何の抵抗もなく大文字のコードによりコード化されている。これも[4]の定義からすると、立派な包摂である。しかし、後にタイプライターのキー数が増えたり文字コードが 7 ビットになったりした時に、これらの包摂が JIS の「高（くち)」と「髙（はしご)」のような問題を起こしたということはなく、異なる文字としてあっさり分離されているし、それに対する異論は聞いたことがない。

3.AD/DA 変換と包摂

ここで AD/DA 変換として、0~1V の範囲の電圧を 2 ビットで線形に表現する場合を考える。00 には 0.125V、01 には 0.375V、10 には 0.625V、11 には 0.875V が対応するが、これを、以後、代表電圧と呼ぶ。このとき偏りのない理想的な AD/DA 変換器を考えると、入出力は図 1 のようになる。

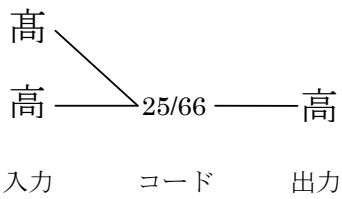

図 2.理想的な文字コード入出力

図 1 で明らかなように、入出力に対称性はない。0.01V も 0.249V も 00 にＡＤ変換されるが、DA 変換においては、00 は代表電圧である 0.125V にしかならず、0.01V や 0.249V が出力されることはない。0.125V は、同じコードに対応する電圧範囲（0V ～ 0.25V）の中央値であり、ビット数がもっとあった場合に本来出力されたであろう電圧との平均誤差が最も小さく偏りも 0 であるという意味で理想的だからである。

図 1 の AD 変換器に 1.5V の電圧が入力された場合は範囲外としてエラーとしてもいいが、0.24V の電圧が入力された場合は、当然 00 にエンコードすることになる。出力は代表電圧の 0.125V となりかなり離れているが、ビット数が少ないことによる必然的量子化誤差である。同様に、代表字体に「高（くち)」を含むが「髙（はしご)」を含まない文字コードで「髙（はしご)」という文字を入力しようという場合は、先の議論で電圧 1.5V というより 0.24V に相当し、エラーにせずに「高（くち)」として入力し代表字体の「高（くち)」として出力するのは当然である（図 2)。

実は活字においても、事情は同じである。「高（くち)」を含み「髙（はしご)」を含まない活字を使う場合、文選工は「髙（はしご)」の字を見たら「高（くち)」の活字を拾い、印刷結果も「高（くち)」となる。[2, 4]で、出力において「髙（はしご)」の字が許容されるというのは、DA 変換の理想的な出力電圧との対比でもおかしいし、活字文化の否定でもある。手書きの場合も、「高（くち)」が基本の字で「高（くち)」と「髙（はしご)」を区別する必要がないと考える人間は、「髙（はしご)」の字を見たら「高（くち)」の字だと思い、筆写する場合も「高（くち)」と書く。

以上の議論により、図形の文字コード化という入力における包摂は、ＡＤ変換の際の量子化誤差と同様の現象でしかないことがわかる。また、DA 変換の際の理想的な出力電圧が代表電圧であるのと同様、文字出力の際の理想的な出力字体は代表字体である。

では、出力における包摂はどう説明できるのだろうか？現実の DA 変換の場合も、00 に対して代表電圧の 0.125V しか出力さ

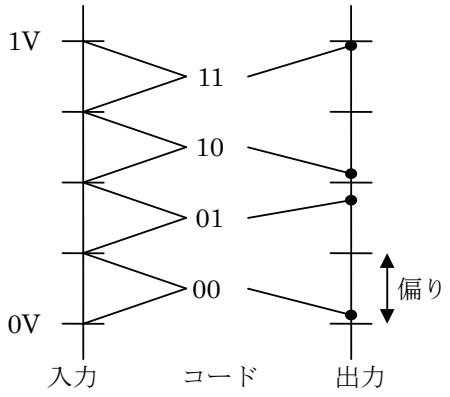

図 3 出力に±1/2LSB の偏りが許される AD/DA 変換

れないわけではない。代表電圧は平均誤差を最小にし偏りをなくすための理想的な電圧だが、工学においては誤差（この場合は偏り）を避けられるものではないからだ。AD/DA 変換では古典的な許容偏りとして±1/2LSB まで認めることが多いが、本稿の場合これは 0.125V に相当し、00 に対して 0.01V や 0.249V が出力されてもこの範囲内である。漢字の場合、[4]の受信装置の適合性規準に「同じ種類の図形文字中の他のいかなる図形文字とも区別できなければならない」とあるのは、±1/2LSB 未満か、より厳しい許容偏りを要求していることに相当する。±1/2LSB 未満の偏りの DA 変換（図３）では、異なるコードが同じ電圧になることがないからである。±1/2LSB の偏りは、丁度量子化誤差と同じ大きさでもある。しかし、より平均誤差や偏りを減らすために±1/4LSB の偏りしか許容しないこともある。

逆に、ビット数の多い AD/DA 変換では、±1/2LSB 以上の偏りが認められるのが普通である。こういう場合単調性も重要であり、全体の偏りや利得を調整した後の個々の偏りは INL（Integral Non-Linearity、最大値と最小値を結ぶ直線（代表電圧）からのずれ、本稿でここまで偏りとしてきた量）と DNL（Differential Non-Linearity、隣接するコード間の差の LSB からのずれ）で指定されることがある。適当にググったところ、例えば 16 ビット ADC で INL が±6LSB、DNL が±2LSB というものは普通にある。これは、隣接するコード間で、理想的には LSB だけの増加しなければいけないところ、[-LSB, 3LSB]の増加が認められることを意味する。つまり、単調性といっても単調増加でなければいけないわけではなく、多少の減少もあっていいのである。16 ビットでも INL、DNL ともに±LSB 以下という変換器もないわけではないが、高価であるようだ。

この話を文字コードにあてはめると、漢字のように似た異体字が多数ある場合、それらを個別に代表字体としても、突然ある異体字が現われたときは他の似た異体字と混同するかもしれない(INL が大)が、似た異体字を直接注意深く比較すれば差が認識できる（DNL が小)ということになる。また、それほど注意しない場合には似た異体字は混同されがち(DNL が大)だが、だからといって代表字体の数を減らす必要はないということでもある。[4]の「同じ種類の図形文字中の他のいかなる図形文字とも区別できなければならない」という規定は、遵守することは不可能ではないが、字種の多い場合には、異なるコードの（入出力偏りとしての意味での）包摂範囲が重なりあうのは当たり前であるということになる。検索においては、似た異体字を区別しないあいまい検索は必須であるが、ラテン大小文字を区別しないあいまい検索と本質的に同じことである。

以上の議論により、文字コードの図形化という出力における包摂は、DA 変換の際の出力偏りと同様の現象でしかないことがわかった。

なお、同じビット数の DA 変換器でも、その偏りにはグレードに応じた違いがあるのと同様、同じ代表字体の文字コードでも実装のグレードに応じて異なる偏りのものが認められてよく、用途によって使い分けられるべきである。ところで、1 ビット DA 変換ともいわれるΔΣ方式による DA 変換では、誤差の一部は量子化誤差の周波数領域でのスペクトラムとして定義され、低周波における誤差は小さくないといけないが、高周波における誤差は大きくてよい。図形のスペクトラム領域で文字の誤差を規定してもあまり意味はないだろうが、同じ文書の中でも、文脈に応じて文字の許容偏りを

変えることには意味があろう。これは、常用漢字が固有名詞の表記には適用されないこと等に相当する。

4.JIS 漢字コードのありかた

前節の議論により、JIS で包摂とされる概念は、入力においてはＡＤ変換では避けられない量子化誤差、出力においては出力の偏りとして整理できた。しかし、これら二つの概念を一括りに包摂としてしまったのが現在の JIS である。

なお、制定経緯等からそうでないことは明らかだが「JIS は出力の許容偏りとして±1/2LSB に相当するものを規定し、これが量子化誤差と同じなので、両者をまとめて包摂と読んでいる」と善意に（というか無理に）解釈することも不可能である。偏りは入力においても必然的に発生するものだが（図 4）、[4]における包摂範囲を厳密に規定しようとの試みは、入力における偏りを±1/2LSB より遥かに小さく、限りなく小さくしようというもので、工業標準としては問題である。なお、図 4 では、コード 01 に対応する出力電圧が、同じコードに対応する入力電圧範囲外にあり、コード 01 の出力をさらに同じ特性の AD 変換器で入力すればコード 10 になってしまうが、誤差というものはそういうふうに累積して大きくなるものであり、目くじらを立ててもしょうがない。「写本の際に字体が変わることなど当たり前である」と説明すれば、漢字学者も納得するだろう。累積誤差を小さくしたければ、出力をなるべく代表字体に近い字体で行うべきである。

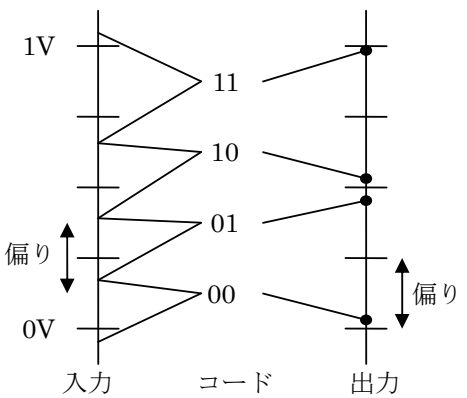

図 4 入出力に±1/2LSB の偏りが許される AD/DA 変換

JIS（に限らず文字コード規格全般）のありかたとしては、平均誤差が最小で偏りのない代表電圧に相当する代表字形の表を定め、入力、出力において許容される偏りを定めるべきである。

量子化誤差は特に指定するまでもなく、[2]のように代表字形の文字表だけを与えれば十分である。普通の文字コード規格は、そうなっている。[4]では、包摂範囲について大量の規定があるが、例外は多く、複数の規定の適用の結果衝突が起きた場合どうするかなどは常識に頼らざるをえず、実はあまり緻密なものではない。一方文化審議会国語分科会は、常用漢字の字体を定めたが、本来の用途である公文書ではその字体しか使わないので、偏りの問題は生じない。

入出力の許容偏りは、文書中でも変動するものであり、入出力装置としてというより、入出力が行われる時と場合に応じて使用者がそれぞれ指定するしかない。使用者が[4]と[5]のような複数の漢字コード規格を組み合わせて字種を増やす場合は、AD/DA 変換のビット数を増やすことに相当する。この際、電圧の範囲を 2V に拡大するために 1 ビット増やすなら量子化誤差は変わらないが、代表電圧の間隔を詰める場合は量子化誤差は小さくなり偏りへの要求は厳しくなるものと予想される。文字数を増やす場合は、[5]のように両方の理由が混在していることも多い。

受信装置の適合性は、実際に出力される字形の文字表により、使用者が判断することとなる。送信装置の適合性は、仮名漢字変換の場合は漢字候補を提示する表示装置の文字表から判断できるが、OCR の場合は仕様書を読解するなり使ってみるなりして判断するしかないだろう。

代表電圧の規準は標準原器などによるが、同様に、代表字体は常用漢字表や康熙字典に基づいて決めればよい。標準原器や宇宙定数はほとんど変動しないが、文化はそれに比較すれば日々発展しており、常用漢字表等の字体が変わった場合に JIS の字体を変えるのは、当然と言えよう。

5. 熱雑音等

工学においては、偏りによる誤差以外に

も熱等による雑音も避けがたい。文字の場合は、文選工の活字の拾い間違い、仮名漢字変換での変換ミス（異体字の認識間違いは偏りであり、別の文字と認識しているものをうっかり混同するのが雑音である）等は熱雑音に相当し、有限のドット数で図形を表現することによる誤差はショット雑音に相当するといえなくはない。ただ、これらの雑音は慎重に入力する（温度を下げる）ことやドット数を増やす（大電流を流す）ことによって減らすことができ、文字コードとして考える必要はないと思われる。

6.Unicode について

以上のように包摂を入力と出力に分けた議論に基づくと、Unicode で日中韓の漢字に同じコードを割り当てている問題は、入力においてはあまり問題にならないが、日中韓で代表字体が異なる以上、出力で正しい代表字体を選択することができないことが問題だと明確になる。個々の字については、さらに、日中韓で許容偏りが大きく異なる字や、ある国の代表字体が他の国の許容偏りの範囲外にある字（「骨」など）があるため、より大きな問題となる。

同じコードにしてしまった漢字を分ける情報をオプションで与えても、オプションでしかなければ省略されることもあり解決にならない。根本的解決は、どれか一国の漢字は今のコードのままとするとしても、他の漢字は代表字体ごとに新たなコードを与えるしかない。

Unicode は、他にも、双方向性の扱いで有限状態性を失っている等の問題もあり、もはや見捨てたほうがいいかもしれない。

7.終わりに

もともとデジタルなものである文字の長い歴史において、JIS 漢字コードで唐突に出現した包摂という概念には、これまで工学的にまともな定義がなかったが、電圧の AD/DA 変換との対比により、JIS における包摂は入力における量子化誤差と出力の偏りで説明できることを示した。

現在の JIS は、包摂を量子化誤差と出力の偏りに分離し、入力の偏りも導入し、実在する AD/DA 変換器との対比により工学的に意味のある形に改版する必要がある。

参考文献

[1] 太田昌孝、「いま日本語が危ない」、ISBN4-89542-146-5、光芒社、1997.

[2] 「情報交換用漢字符号系」、JIS C 6226-1978、1978.

[3] “Universal Multiple-Octet Coded Character Set (UCS) - Part 1: Architecture and Basic Multilingual Plain”, ISO/IEC 1046-1, 1993.

[4]「7 ビット及び 8 ビットの 2 バイト情報交換用符号化漢字集合」、JIS X 0208:1997、1997.

[5]「7 ビット及び 8 ビットの 2 バイト情報交換用符号化拡張漢字集合」、JIS X 0213:2000、2000.