音声認識の良否の判定基準

6G－3

浦谷 則好+　　鷹見 淳一++
ＡＴＲ音声翻訳通信研究所

１．はじめに

音声認識システムや手法の良否は音声認識率で評価されるのが通例となっている。これは、検索システムの良否を適合率だけで評価することと同じで妥当なことだとは思えない。なぜなら、例えば１０個の単語を認識する音声認識システムを製作した場合、もしこの単語のどれか１つが非常に認識しにくいものであっても、他の単語が完全に認識できれば音声認識率は９０％を越える。しかし、当該の単語を入力しようとすれば何度も発声を繰り返さなければならず、誰もこんなシステムを使いたいとは思わないに違いない。

一般に検索システムは適合率（検索されたデータの中でユーザの要求に合致するデータの比率）と再現率（ユーザの要求に合致するデータの中で検索されたデータ中に含まれる比率）で評価されるのが普通である。もちろん、両者とも高いシステムが望まれるが、適合率と再現率は往々にしてトレードオフの関係になり、検索システム開発者は両者のバランスを図るようにするのが通例である。

われわれは、音声認識システムや手法に対する評価尺度として音声認識率とは異なる新しいものを提案する。そして、検索システムと同様、音声認識率とここで提案する尺度の２つで音声認識システムの良否を判断するようにすることを提案したい。

２．音声入力率の定義

単純化のために単語認識の音声認識システムを例にとる。文認識あるいは音韻認識などの場合も容易に拡張できるので、一般性を失わない。認識対象の単語の集合が$W=\{w_1, w_2, \ldots, w_n\}$のとき、個々の単語の認識率を$p_i$、出現頻度を$f_i$とすれば、全体の音声認識率Pは

$$P=\sum_{i=1}^{n} f_i p_i \Big/ \sum_{i=1}^{n} f_i$$

で求めることができる。

一方、この認識システムを用いて、ある単語w_iを入力する場合の入力試行回数の期待値k_iは

$$k_i=\sum_{j=1}^{\infty} j p_i (1-p_i)^{j-1} = 1/p_i$$

となる。したがって、平均入力回数Nは

$$N=\sum_{i=1}^{n} (f_i/p_i) \Big/ \sum_{i=1}^{n} f_i$$

となる。（ここで、f_iはw_iの入力する要求が生ずる頻度である。）

Nは当然１以上の値を取るので音声認識率Pと一緒に使うためには、あまり望ましくない。そこで、Nの逆数をとってPと値域を同じになるようにする。すなわち、

$$Q=\sum_{i=1}^{n} f_i \Big/ \sum_{i=1}^{n} (f_i/p_i)$$

このQは平均的に（音声）入力の成功する率を示していることになる。そこで、我々はこれを音声入力率と呼ぶことにする。（実はこの式は結局p_iの調和平均そのものである。）

３．音声入力率の計算例

音声入力率の有効性を実証するには例をもって示すのが早道である。そこで、（平均）音声認識率が同じ（0.9）で音声入力率が異なる４つの例をあげる。

ただし、単純化のためf_iは全て同じであると仮定する。またn=10とする。

＜例１＞　$p_1=p_2=\ldots=p_{10}=0.9$ の場合
　　Q= 0.9

＜例２＞　$p_1=p_2=\ldots=p_9=1$ で $p_{10}=0$ の場合
　　Q= 0

＜例３＞　$p_1=p_2=\ldots=p_5=0.95$ で
　　$p_6=p_7=\ldots=p_{10}=0.85$ の場合
　　Q= 0.8972

+現ＮＨＫ放送技術研究所　++現日本ビクター株式会社

An Evaluation Criterion for Speech Recognition

Noriyoshi URATANI and Jun'ichi TAKAMI
ATR Interpreting Telecommunications Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan

＜例４＞ $p_1=p_2=...=p_8=0.98$ で
$p_9=0.90$, $p_{10}=0.26$ の場合
Q= 0.7622

（定義式からも明らかだが、）これらの例から音声入力率は必ず音声認識率よりも大きくはならないことが分かる。２つが同値をとる場合は個々の（単語の）音声認識率が全て等しい場合に限られる。上の例からは、音声翻訳システムのユーザの立場に立てば音声認識率よりもむしろ音声入力率の方が実感に近いと想像できる。例えば、１つでも音声認識率０のものがあれば音声入力率は必ず0を示し、そんなシステムは使いものにならないことが分かる。また、例３、例４からは個々の音声認識率の分散が小さければ、（全体の）音声認識率と音声入力率の差は小さいことも見てとれる。

以上見てきたように、ここで提案する音声入力率は音声認識率の基準としての不完全性を補完する有用な基準である。

４．音声認識率と音声入力率のトレードオフ

先に述べた＜例１＞のシステムはどんな入力をいれてもw_{10}とは認識しないシステムであることが分かる。そのかわりw_1からw_9に対応した入力に対しては必ず正しく認識するものである。したがって、もしシステムがw_1からw_9の９個の認識で良いならば完璧な音声認識システムだと言える。しかし、１０種の認識を要求される以上このままでは音声入力率（=0）が示すように使いものにならないことが確かである。

一例として、w_{10}に対応する入力に対しては、認識結果はw_1からw_5に等しく分布し、w_6からw_9に認識されることはないと仮定する。（単純化のため入力要求の頻度はw_1からw_{10}まで全て同一とする。）つまり、w_1（w_2, w_3, w_4, w_5）と認識される入力の内、12回に２回はw_{10}に対するものであることが分かる。そこで、w_1（w_2, w_3, w_4, w_5）と認識する内の2/12を（無差別に）w_{10}と（認識）出力することにする。そうすれば、w_1（w_2, w_3, w_4, w_5）に対する認識率は10/12(0.833)に低下してしまうが、w_{10}に対する認識率は2/12(0.167)に向上する。したがって、このような変更を加えたときのシステム全体の音声認識率Pと音声入力率Qは

P=(10/12*5+1*4+2/12)/10=10/12
Q=10/(12/10*5+1*4+12/2)=10/16

となる。すなわち変更によって音声認識率は0.9から0.833に低下し、反対に音声入力率0から0.625に向上したことになる。実際に音声認識システムを製作する場合にも似たような局面が予想される。極端に認識率の悪い単語（あるいは文、音韻など）があったのではシステムとしては使いものにはならない。そこで、全体の音声認識率の多少の低下を招いても認識率の悪い単語の認識率の向上に努めるのが当然である。このことが、結局ここで提案した音声入力率の向上に結びついていることが分かる。すなわち、一般に音声認識率と音声入力率とのトレードオフを図らなければならない。（この例の場合、上述した方法はQを最大にする仕方になっていない。これを求めるにはw_1（w_2, w_3, w_4, w_5）と認識する内、w_{10}と（認識）出力する確率をpとおいて、Qをpの式で表してその最大値を求めればよい。この例では$p=(5-\sqrt{5})/4$(=0.6910)のとき最大でQの値もpと等しくなる。）

上の例では誤りを非常に単純なものに仮定したが、実際のシステムでもし誤りの傾向についてもっと情報が得られるのならば「音声認識率の低下は少なく、音声入力率の増加は大きく」なるような変更が可能であろう。

５．おわりに

音声認識システム（あるいは手法）の良否の判定基準として新しい基準（音声入力率）を提案した。例を用いてこれが音声認識率よりユーザの実感に近いものであることを示した。また、音声認識率と音声入力率はトレードオフの関係にあることも示し、２つを調整する方法についても述べた。われわれは今後、この基準が音声認識システムの判定基準として（平均）音声認識率と一緒に使われることを願っている。

参考文献

中川聖一：確率モデルによる音声認識，電子情報通信学会編，コロナ社，1988