G-27

有声音／無声音情報を用いた単語認識

Spoken word recognition using Voiced/Unvoiced information

○金澤　常助
Tsunesuke Kanezawa [1]

Abstract: One way to recognize the pattern of spoken word is using Mel-Frequency Cepstral Coefficients (MFCC). This paper proposes the way to recognize the patterns using only an one-dimensional vector. The vector contains the information of Voiced/Unvoiced through the wave, from the twelve-dimensional MFCC data. The simulation results show that such a manner of recognition is viable in some cases.

１．概要

音声認識では多次元の特徴量で構成される Mel-Frequency Cepstral Coefficients （MFCC）を用いて行う方法があるが，本論文では、その単一次元の特徴量のみを用いて単語認識を行う方法を提案する．また、シミュレーション実験を通して単語パターン識別の方法としてある程度有用であることを実際に確認した．

２．MFCC データを元にした単一次元の特徴抽出

２.１．MFCC

一般に、音声認識において、原波形として得られた音声サンプルをそのまま用いるのではなく，特徴量として MFCC が用いられることが多い．この MFCC 特徴量は一つの音声サンプルに対して、一つのフレームにつき n 次元の特徴量を持つような複数フレームで構成され，口腔形状を元にした音響特性のみを抽出したものである．

また、次元数 n が 12 であれば音声認識を行うことは十分可能であり、特に個人差のあるピッチの情報は含まれないものであるので識別する際に用いる特徴量として適している、ということが知られている．

２．２．First MFCC

MFCC の 12 次元の特徴量のうち第 1 次元目の値（First MFCC）の数値のみを用いることにより有声音および無声音の識別が可能になるという報告がある[1]．また、その値は有声音であれば大きくなり，無声音であれば小さくなるということも知られている．

つまり、この数値は True/False の２値ではなく増減を含むような連続値なので，その数値の変化を考察することにより，さらに細かい識別が可能であると考えられる．

３．実験

本研究では、数字の音声 1/ichi/, 2/ni/, 3/san/, 4/yon/, 5/go/，6/roku/，7/nana/，8/hachi/，9/kyu/を用いて認識実験を行った．学習データとして各数字につき 10 個ずつ，また、同数（10 個）のテストデータを準備した。各音声サンプルは 48kHz，16 ビット，モノラルで録音し，その後、MFCC に変換して実験を行った．

次に、全ての学習データについて得られる First MFCC（$allfm$）から有声音／無声音の客観的基準として以下(1)式を用いて$base$を導出した：

$$base = \frac{max(allfm) + min\ (allfm)}{2} \tag{1}$$

さらに、上式を利用して以下(2)式によって示されるように、1 学習データにつき、各 First MFCC（fm）から、有声音／無声音のみを示す特徴量としてベクトル S を生成した．

$$fmのi番目の要素:fm_i$$

$$S_i = \begin{cases} 1\ (fm_i > base) \\ 0\ (fm_i < base) \end{cases} \tag{2}$$

また、サンプルデータごとに波形データの時間的違い，すなわちフレーム数には違いがあったため，扱いやすさおよび客観的考察を行うためにすべての S に拡縮を行い 10 列に置き換えた．

(1)式を用いて変換すると今回$base = -3.95205$であったため，以下例としてfmを Figure 1.が示す 6_1.mfc とすれば，実線部$base$を境目にして有声／無声／有声の順になっていることが確認できる．これより得られた S は

$$S = [1\ 1\ 1\ 0\ 0\ 1\ 1\ 1\ 1\ 1\ ] \tag{3}$$

となる．

次に Figure2.は/nana/の First MFCC の一例を示している．この例ではfmが常に$base$を上回っているため，7_4.mfc における S は

１：日大理工・学部・数学　２：日大理工・教員・数学

$$S = [1\,1\,1\,1\,1\,1\,1\,1\,1\,1\,1] \quad (4)$$

となるが，その中でも値の増減は確認できる．この情報をさらに特徴量として追加して認識を行う．実験を通して確認された増減情報は全部で 12 パターンあり，この情報をそれぞれ，6 列の 1，0 の組み合わせで表現されるものに対応させた．これをudと呼ぶことにする．Figure2.が示す 7_4.mfc の増減は増/減/増/減である．7_4.mfc におけるudは

$$ud = [1\,1\,1\,0\,0\,0] \quad (5)$$

となる．

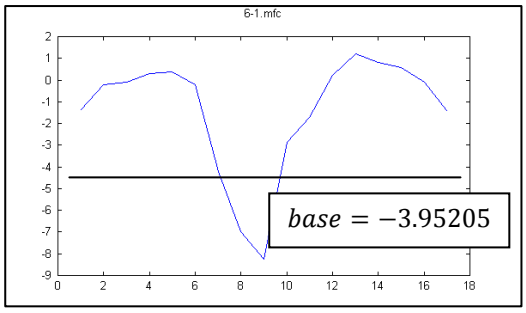

Figure1. fm of 6-1.mfc

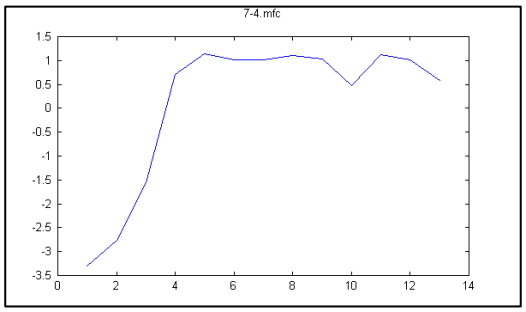

Figure2. fm of 7-4.mfc

これらSとudを並べて作成した(6)式に示す特徴ベクトルHを用いて学習データから距離の近似により識別辞書を作成し，テストを行った．

$$H = [h_1 h_2 \cdots h_{16}] \quad \begin{cases} h_1 \sim h_{10} = S \\ h_{11} \sim h_{16} = ud \end{cases} \quad (6)$$

４．実験結果

前節の実験結果の一部を示す. 以下 Figure3.は各音のテストの成功率である．

結果として/ni/, /san/, /hachi/, /kyu/の識別は可能であることが確かめられた．/ichi/, /yon/ , /go/, /roku/ , /nana/に関しては今後改善が必要である．

また/yon/にテストを行うと/nana/または/kyu/として認識される. いずれの結果も 50%の確率で/nana/または/kyu/となってしまった．

有声音／無声音の組み合わせおよび子音の構成が近いと識別は困難である. 特に困難な例を Table1.に示す．

今回用意したサンプルにおいて/ni/と/go/の違いを識別するのにこの First MFCC だけでは 2 サンプル間の違いがほとんど見られない為，最も困難な課題であると考えられる．

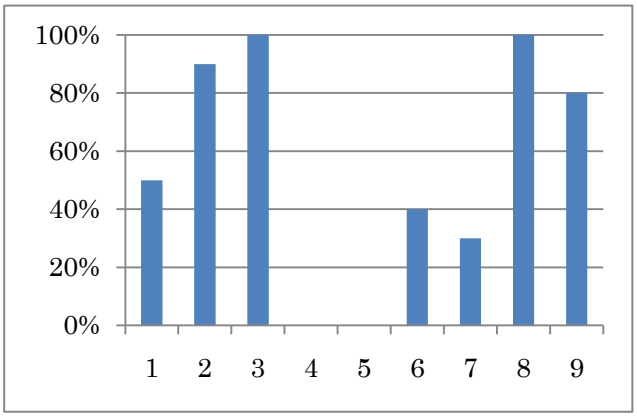

Figure3. Recognition result

テストの結果が期待した結果と異なっていたため，/ni/と/go/についての学習データ同士および/nana/と/ni/に対して識別を行い検証した結果はそれぞれ 92.94%, 94.82%で一致してしまった．

これは識別に用いる特徴量として First MFCC の増減情報よりも有声音／無声音情報に比重が寄っている為であると考えられる．

Table1. The rate of the recognition (unexpected)

H_1	H_2	計算量	合致数	合致率(%)
/ni/	/go/	1700	1580	92.94
/ni/	/nana/	1700	1612	94.82

従来の 12 次元での識別に対して，単一次元で識別した結果として現段階でのスコアは不完全ではあるが，識別精度を上げることは可能である．

したがって今後の課題は細部にわたり実験を見直し，より精度を上げることである．あくまで認識に用いるデータは First MFCC のみとし，どのレベルまで単語認識が可能なのかを検証することである．

５．参考文献

[1] Jashmin K.Shah, Ananth N.Iyer, Brett Y.Smolenski, and Robert E.Yantorno, “ROBUST VOICED/UNVOICED CLASSIFICATION USING NOVEL FEATURES AND GAUSSIAN MIXTURE MODEL”, IEEE ICASSP’04, 2004.

[2] Tetsuya Hoya and Cees van Leeuwen, “A cascaded neuro-computational model for spoken word recognition”, Connection Science, Vol.22,No.1,pp87-101,2010.