X-Yトラッカを用いた口形情報併用による単語音声認識

古谷 忠義*・添田 満*・黒須 顕二**・田村 進一***

Speech Recognition with Lip Movement Data Using an X-Y Tracker

Tadayoshi FURUYA*, Mitsuru SOEDA*, Kenji KUROSU** and Shinichi TAMURA***

Either noises or similar word syllables deteriorate the speech recognition when the methods use only voice or sound data. Additional information applied in lip-reading may improve the recognition under such conditions.

This paper has proposed a system of speech recognition for some finite number of words under noisy background. The system, which has both an X-Y tracker and a microphone, can detect sounds and positions of lip movement. The X-Y tracker processes the visual images and sends a pair of the X-Y positions in the coordinate, which enables us to obtain the lip movements in real time without some complicated tasks to the main computer. The pair of the marked position near the lip in addition to the speech are inputted to form a pattern for speech recognition. The patterns of the voice synchronized with the X-Y positions of the lip movements are compared to the standard patterns to recognize the spoken words. The weight coefficients to change the importance of visual or voice data are introduced and selected to get the best performance for some registered words.

The experiments were carried out to prove that the proposed system can improve the recognition rate in the presence of continuous noises. The results has shown about ten percent improvement under signal noise ratio as 35 dB, and about 20 percent under the signal noise ratio as 26 dB.

Key Words: speech recognition, lip reading, DP matching, noise, X-Y tracker

1. はじめに

機械の操作や運転では手足などを使うことが一般的であるが、これらの通常の人間の操作機能がふさがったときには音声による補助入力手段になる。特に人間の限界に近い高度な操作の必要となる複雑な操縦では音声による指示を与えることができれば、操縦者の負担が大いに軽減できる。これらのための限定単語の音声認識はすでに実用段階にあるが、騒音の激しい環境のもとや母音の類似した単語が存在するとき、音声のみに頼る従来の音声認識法では認識率の向上がなくなる可能性がある。

騒音環境下における音声認識に関しては騒音の定義が難しいうえに騒音で音声が消される状態になれば認識できなくなることは自明であるので騒音に対する認識率の研究は、雑音のある環境下で有効に音声を取り出す方法や¹⁾、雑音のキャンセリング、スペクトルのサブトラクションという観点から認識率の向上の研究が行われている^{2),3)}。これらの雑音の存在する場合、音声を有効に取り出すことはもちろん、除去することのできる雑音は可能な限り取り除く必要がある。しかし実用になる程度まで雑音の影響を軽減できないときは他の方法を併用しなければならない。

一方口形情報を利用して認識率をあげることに関連しては読唇に関する研究が行われており、聴覚障害者のために読唇、読唇の練習をするソフトの開発や画像処理などを用いて読唇の実現をめざす研究が行われている^{4)~7)}。筆者らも画像処理を利用し、母音口形の識別と音声認識の研究を行っている^{8),9)}。さらにニューラルネットを用いて音声と口形特徴量を併用する母音の認識についても研究されている^{10),11)}。これらの読唇の研究では口形情報処理の高速化の問題と音声処理を組み合わせる単語認識率の向上の可能性などを検討す

* 北九州工業高等専門学校 北九州市小倉南区志井 5-20-1

** 九州工業大学工学部 北九州市戸畑区仙水町 1-1

*** 大阪大学医学部 大阪市北区中之島 4-3-57

* Kitakyushu College of Technology, Kitakyushu

** Faculty of Engineering, Kyushu Institute of Technology, Kitakyushu

*** Osaka University Medical School, Osaka
(Received December 25, 1990)

(Revised March 13, 1991)

る必要がある。

本論文は、X-Yトラッカを用いて唇の一点の位置を測定、処理することによる高速化および音声処理と口形情報の処理の併用による暗音環境下の単語認識率の向上の可能性について検討したものである。

高速化については、X-Yトラッカが画像処理によるマークされた1点のXY座標値をデジタル出力してくれるので本体のコンピュータはそのデータを受けとるだけで画像処理をしなくてよいという利点がある。また、音声処理と併用することによって口形情報との時間のマッチングがとりやすく、口形の無意味な動きによる誤動作や騒音下における音声認識の困難さを克服できるものと考えられる。

これらの利点と可能性を示すために構成されたシステムと実験例について説明する。

2. 装置の構成

口形情報と音声をコンピュータに取り込み単語認識を行うために Fig. 1 に示すシステムを構成した。中央演算装置として32ビットのミニ・コンピュータ(三菱 MX-3000)を採用し、周辺機器を接続している。

音声信号は、被験者の口の近くにマイクロホンを取り付け、アンプで増幅した後 A/D 変換されてコンピュータに取り込まれる。A/D 変換器はサンプリング周波数を 12kHz に選び、およそ音声の全振幅を 12ビットで量子化するように設定した。

A/D 変換器はコンピュータ MX-3000 本体に内蔵のもので A/D 変換速度 10 μ s, DMA 転送できるものである。

口唇の動きを実時間で計測するために X-Y トラッカ (ARTIST-G 120) を用いている。これはカメラ一体型の画像処理装置で画像内の対象の特徴のある対象点の位置や対象の2点の相対位置、距離、角度の計測を 60 Hz のサンプル周波数で計測するものである。

Fig. 1 Speech recognition system

対象の点が動く場合も画面上の点を自動追跡し、ある瞬間の対象の位置を計測することができる。計測された結果は RS 232 C インタフェースを介してコンピュータに送られる。

ここで口唇の動きをとらえるために、白色のマーカを下唇の先端につけ、顔面が動いても下唇だけの動きがとらえられるように固定点として鼻の先端にも白色マーカをつけた。固定点を原点とする下唇のマーカの相対位置を約 45 度前方から X-Y トラッカで計測し、これを口形情報とした。なお顎の動き、唇の突き出し、口の開口面積などの情報が音声認識に有効なことは確かめられている⁸⁾。

音声および X-Y トラッカによる相対位置は実時間で計測され、コンピュータのメモリーに現時点から 5 秒前までのデータが常時格納されているように動作するようプログラムされている。

また実時間処理しているので画像は音声と同期して取り込まれるように構成されている。

単語の認識を行うときには記憶されているデータから心要部分を取り出し画像と音声を併用した認識が行われる。

3. 口唇の動きを併用する音声認識法

3.1 信号処理の過程

ここで構成したシステムはよく用いられる音声による単語認識法に口唇の動きの情報を加えたものである。

その信号処理の流れを Fig. 2 に示す。音声の信号の 2 乗平均値があるいき値を超えるかあるいは口唇の動きが開始されたと判断されることによってデータの切り出しが開始される。音声と口唇の動きは常時データとして貯えられているのでデータの開始点は現時点より以前にとることができる。ここで音声の 2 乗平均値が大きくなるのは発音開始の少し前であるのでデータの開始点はいき値を超える時間より 0.3 秒前をとっている。

取り込まれた音声のデータと X-Y トラッカによる口唇の動きはパターン作製のために多数のフレームに分割され、各フレームごとに特徴データに変換され格納される。音声のサンプル周波数を 12kHz に選んだときのフレームの構成法を Fig. 3 に示す。まず約 2 秒間のデータが 60 個のフレームに分割される。1 フレームは 4 時点の口唇の動きのデータと 606 個 (4 kHz のサンプル周波数のときは 202 個) の音声原データからなり、時間の長さは 3/60 s である。各フレームは 1/60 s の重なりを前後のフレームともつように

Fig. 2 Diagram of signal processing

Fig. 3 The composed data frames

構成した。これは重なりあう部分をとることによってデータの減少を防ぐためである。

606 個の音声データは以下に示す手順によって音声の特徴パターンとなり同じ 60 フレームのパターンとして登録される。

一方口唇の位置の特徴量として X-Y トラッカにより計測された 2 点の位置から口唇の下位置の相対位置が登録される。

3.2 特徴パターンの作成

取り込まれた音声データと口唇の動きのデータから各フレームごとに特徴パターンを作成する^{(12), (16)~(19)}。

音声データについては LPC スペクトル分析を行う。まず、得られたデータ $\{x_i\}$ から、次式により $i=1, 2, \dots, 12$ に対する自己相関関数を計算する。

$$r_i = \frac{1}{r_0(N-i)} \sum_{j=1}^{N-i} (x_j - \bar{x})(x_{j+i} - \bar{x}) \quad (1)$$

ここで $\bar{x}$ は平均値、 N はデータ個数である。

これらの相関関数から、つぎのレビンソンダービンのアルゴリズムにより線形予測係数 $\{a_i\}$ を求める。

$$\begin{aligned} a_{p+1}(p+1) &= -\frac{1}{\sigma(p)^2} \left\{ r_{p+1} + \sum_{i=1}^p a_i(p) r_{p+1-i} \right\} \\ a_k(p+1) &= a_k(p) + a_{p+1}(p+1) a_{p+1-k}(p) \\ &\quad (k=1, \dots, p) \quad (2) \end{aligned}$$

$$\sigma(p+1)^2 = \sigma(p)^2 \{1 - a_{p+1}(p+1)^2\}$$

ただし、 $a_i(0)=0$, $\sigma(0)^2=r_0$

ここでは、 $p=1, 2, \dots, 11$ に対し計算を行い、次数 12 の線形予測係数 $\{a_i(12)\}$ ($i=1, 2, \dots, 12$) を求めた。

線形予測係数 $\{a_i\}$ ($i=1, 2, \dots, q$) が得られると、LPC ケプストラム係数 c_n は、

$$c_n = -a_n - \sum_{i=1}^{n-1} \left(1 - \frac{i}{n}\right) a_i c_{n-i} \quad (1 \leq n \leq q) \quad (3)$$

$$c_n = -\sum_{i=1}^p \left(1 - \frac{i}{n}\right) a_i c_{n-i} \quad (n > q) \quad (4)$$

ただし、 $c_0 = \log \sigma^2$, $c_1 = -a_1$

により求められる。

そこで(2)式により得られた 12 次の線形予測係数を用いて、(3)、(4)式により、24 次の LPC ケプストラム係数 $c_1, c_2, \dots, c_{24}$ を求める。

また線形予測係数を用いて $j=13, 14, \dots, 24$ に対する自己相関関数 $r_{12}, \dots, r_{24}$ を次式により求める。

$$r_j = -\sum_{i=1}^{12} a_i(12) r_{j-i} \quad (5)$$

こうして得られた自己相関関数と LPC ケプストラム係数を用いて、時刻 t (1 フレーム) における区間の音声の特徴をパターンを 50 次元ベクトル

$$X_t = [r_0, \dots, r_{24}, c_0, \dots, c_{24}]$$

で表現し、一単語の音声から、フレーム数分の特徴パターンを作成する。

口唇の動きについては、1/60 秒ごとに、下唇につけたマーカと鼻の先端につけた固定マーカとの相対座標 $(x(t), y(t))$ が得られるので、1 フレームにおける区間の口形特徴パターンをこれらの 4 個の座標をならべた 8 次元ベクトル

$$X_p = [x_1, y_1, \dots, x_4, y_4]$$

で表現する。

この X_t と X_p を一つにまとめて

$$X_q = [X_t, X_p]$$

$$=[r_0, \dots, r_{24}, c_0, \dots, c_{24}, x_1, y_1, \dots, x_4, y_4] \\ q=1, 2, \dots, 60 \quad (6)$$

とし、この Xq を各フレーム（時刻）の総合特徴パターンとする。一単語の入力データから、これらの特徴パターンをフレーム数分作成する。

3.3 特徴パターンの距離尺度

各フレームにおける入力音声と標準音声の特徴パターンの距離尺度には、WLR 尺度を用いる¹⁶⁾。これは、ホルマントのようなスペクトルの山部のずれに対してセンシティブになるようにしたもので

$$d_o(f, g) = \sum_{i=1}^{24} (r_i^{(f)} - r_i^{(g)})(c_i^{(f)} - c_i^{(g)}) \quad (7)$$

により計算される。ここで r_i , c_i は自己相関関数および、LPC ケプストラム係数で、 f , g はそれぞれ登録してある標準データと入力データを表わす。

また、口唇の動きの標準パターンと入力データの距離尺度は、

$$d_m(f, g) = \sum_{i=1}^4 (|x_i^{(f)} - x_i^{(g)}| + \alpha |y_i^{(f)} - y_i^{(g)}|) \quad (8)$$

と定義する。ここで x , y は唇と鼻につけられたマーカの相対座標である。また α は、口唇の左右、上下の動きをそれぞれの程度重視するかの重みである。

総合的な単語認識のための距離尺度としては、これらの音声特徴と口形特徴を組み合わせたものを考える。すなわち、各フレームの標準データと入力データのパターンの距離をつぎのように定義する。

$$d_i(f, g) = \beta d_o(f, g) + (1 - \beta) d_m(f, g) \quad (9)$$

ここで β は、認識において音声と口唇の動きの特徴をそれぞれの程度重視するかの重みを表わす。

3.4 パターンマッチング

単語全体を標準パターンとする認識システムでは、標準データと入力データの二つのパターン間の全体の距離を定義する必要がある。この場合、問題となるのは発声速度による音声パターンの変化である。発声速度が変化するとき、単語内の母音の定常部の長さの変化は、子音やわたりの部分の変化に比べて大きい。このため、標準パターンの音韻と入力音声の音韻とがうまく対応するように時間軸を非線形に伸縮させる必要がある。

そこで本研究における認識システムでは、DP マッチングにより時間軸整合を行い、標準パターンと入力パターンの距離を求める。

いま対応すべき二つのパターンの特徴ベクトルの系列をつぎのように表わす。

標準パターン $F = f_1, f_2, \dots, f_j, \dots, f_J$

入力パターン $G = g_1, g_2, \dots, g_i, \dots, g_I$

ここで f , g はそれぞれ (6) 式で表わされる入力データ、標準データの各フレームの特徴パターンベクトルである。

標準パターン F の時間軸 j を入力パターン G の時間軸 i に写像する関数を $j = j(i)$ とする。これにより、伸縮されたパターン $F' = f_{j(1)}, f_{j(2)}, \dots, f_{j(I)}$ と入力パターン G との距離を最小にするように、つぎのような距離を定義する。

$$D(F, G) = \min_{j=j(i)} \left[\sum_{i=1}^I d(i, j) \right] \quad (10)$$

$d(i, j)$ は二つのパターンのベクトル g_i と f_j の距離を表わし、実際には (7) ~ (9) 式により計算される。

いま、 G と F の始端と終端はそれぞれ一致するから $j(1)=1$ かつ $j(I)=J$ となる。また発声速度が変わっても音韻生起の順序は変わらないから、 $j(i)$ は単調増加関数である。これらの条件のもとで、(10) 式の最小化の DP アルゴリズムがつぎのように得られる¹⁰⁾。

初期条件 $g(1, 1) = d(1, 1)$

$$\text{漸化式 } g(i, j) = d(i, j) + \min \begin{cases} g(i-1, j) \\ g(i-1, j-1) \\ g(i-1, j-2) \end{cases}$$

この式で $j=J$ になるまで i を変えながら計算していくと $g(i, j)$ が求まり、これから時間正規化後の距離は

$$D(F, G) = g(I, J)$$

となる。

このようにして、入力音声の特徴パターンとすべての事前登録された単語の標準パターンの距離 $D(F, G)$ を求め、その距離が最小となる登録単語を認識結果として出力する。

4. 単語認識実験

本論文で示した手法の評価を行うため、つぎの発音の似た 10 単語

KAMA KAMI KAMU KAME KAMO

KANA KANI KANU KANE KANO (11)

による特定話者の雑音環境下における離散単語認識実験を行った。

4.1 実験方法

Fig. 1 の装置において、音声信号はマイク、アンプを通し、約 1,000 倍に増幅される。アンプの出力側の信号は $\pm 5V$ の制限をかけ A/D コンバータに入力される。口唇の動きは、その位置を話者の前方右 45 度の位置に設定した X-Y トラッカにより計測し、RS-232C インタフェースを介してミニコンピュータ (MX-3000) に入力する。

こうしてまず標準となる単語を入力し、各単語の標準パターンを作成し登録しておく。つぎに、認識実験を行うため、雑音環境下のデータを取り込む。環境雑音は、工場現場などに多い回転機械類の雑音を取り上げた。ここでは、コンピュータのファンの雑音を、マイク、アンプを通しコンピュータのメモリーに格納した。そして雑音のない状態で単語音声入力を行い、音声信号が入力されるとコンピュータ内で、メモリーに格納してある雑音データを、その大きさを調整して音声データに加算し、以後そのデータを音声情報が損なわれるという仮定のもとでの入力音声データとして使用した。そしてこれらの入力データから、上述の方法により単語認識を行った。

4.2 重みの決定

事前に登録した各単語の特徴パターンを作成する。上述の方法により、これらの標準データの、パターン I とパターン J ($1 \leq I, J \leq 10$) の DP マッチングを行い、各パターン間の距離 $D(I, J)$ を求め次式を計算する。

$$W = \sum_{I=1}^{10} \left[\min_{\substack{J=1, 2, \dots, 10 \\ J \neq I}} \{D(I, J)\} \right] \quad (12)$$

ここでは単語の違いがより大きな距離として表われるような重みを採用する。すなわち(8), (9)式において、重み α , β を変化させそれぞれの重みにおける W を求め、その中から W が最大となるときの重みを、この単語認識システムの重みとして採用する。

(8)式の口唇の動きの距離尺度における重み α と(12)式の W の関係は Fig. 4(a)に示すようになった。なおグラフ上、縦軸はそれぞれの W を W の最大値 $W_{\max}$ で割り正規化して表わしている。

この結果から、より単語の違いが大きくなる重みとして、本論文においては、重み α は $\alpha=0.76$ を採用することとした。

同様に(9)式の距離尺度において、音声特徴と口形特徴の重み β を変化させ、それぞれ W を計算した結果、Fig. 4(b)に示すようになった。この結果から重み β は $\beta=0.6$ を採用した。

4.3 実験結果

特定話者の限定単語について認識実験を行った。一般成人男性(学生)を話者として、一つの単語につきそれぞれ50回の認識実験を行った。ここでは、(7)式の音声情報の距離尺度のみによる単語認識((9)式において重み $\beta=1$ の場合)と、口唇の動きの情報も加えた尺度による単語認識(重み $\beta=0.6$ の場合)を行っている。

Fig. 4 Distance vs. weighting parameters

入力信号をサンプリング周波数 12 kHz で取り込んだときの認識率を Table 1 に示す。S/N 比 60 ± 5 dB というのは実験を行った環境の暗騒音のレベルであり、モータ音を採択したものを重量させていない。誤認単語でわかるように類似音のため認識率が 82% とあまり良くない。しかし口形情報を加えることにより、わずかであるが認識率が向上している。

このデータに雑音を重畳し、S/N 比を劣化させ、 25 ± 2 dB に(2人の間の会話がほとんど聞こえない程度)設定した場合、平均認識率は 59% と極端に悪くなる。この場合も口形情報を加えることによって認識率は 83% と大幅に向上することができる。

つぎに子音の認識に必要な高周波の情報が失われ、音声情報が劣化したときを想定し、サンプリング周波数を 4 kHz にしたときの結果が Table 2 である。S/N 比を劣化させ、サンプリング周波数を低くしても、口形情報を加えることによって認識率が極端に悪化されることのないことがわかる。

この実験では(KAMI, KANI), (KAMU, KANU), (KAME, KANE), (KAMO, KANO) など、発音が近い単語の認識は、強い雑音に乱され、認識がほとんどできない状態になっている。母音の部分はパワーが大きく雑音によりあまり乱されないのに対し、子音の部分は雑音に乱されてしまうため、このような結果になるものと考えられる。これに対し、口形情報を加えて認識を行った場合は、その認識率の向上がはっきり見

Table 1 Recognition rates (sampling frequency 12 kHz)

Test word	S/N 60±5 dB			S/N 25±2 dB		
	Only voice	Both image and voice	Mistaken word	Only voice	Both image and voice	Mistaken word
KAMA	80%	98%	KANA	60%	90%	KANA
KAMI	60%	98%	KANI	30%	80%	KANI
KAMU	98%	96%		98%	94%	
KAME	70%	70%	KANE	50%	66%	KANE
KAMO	98%	98%		60%	80%	KANO
KANA	96%	96%		50%	84%	KAMA
KANI	40%	60%	KAMI	50%	76%	KAMI
KANU	86%	90%	KAMU	70%	86%	KAMU
KANE	96%	96%		80%	90%	KAME
KANO	98%	98%		40%	80%	KAMO
Mean	82%	90%		59%	83%	

Table 2 Recognition rates (sampling frequency 4 kHz)

Test word	S/N 35±5 dB			S/N 26±5 dB		
	Only voice	Both image and voice	Mistaken word	Only voice	Both image and voice	Mistaken word
KAMA	92%	90%	KANA	70%	86%	KANA
KAMI	64%	70%	KANI KAMU	34%	64%	KANI KAMU
KAMU	80%	94%	KANU	44%	82%	KANU
KAME	24%	62%	KANE	26%	66%	KANE KANA
KAMO	88%	84%	KANO	72%	80%	KANO
KANA	96%	88%	KAMA	94%	90%	KAMA
KANI	40%	66%	KAMI KANU	28%	60%	KAMI KANU
KANU	96%	94%	KAMU	98%	92%	KAMU
KANE	72%	80%	KAME	62%	78%	KAME
KANO	34%	76%	KAMO	12%	68%	KAMO
Mean	68%	79%		54%	76%	

Table 3 Recognition rates (sampling frequency 4 kHz, S/N=35 dB)

Test word	Only voice	Both image and voice	Test word	Only voice	Both image and voice
ZERO	100%	100%	REI	70%	100%
ITI	96%	100%	ITI	96%	96%
NI	100%	100%	NI	80%	100%
SAN	100%	100%	SAN	100%	100%
YON	98%	100%	SI	64%	100%
GO	100%	100%	GO	100%	100%
ROKU	100%	100%	ROKU	100%	100%
NANA	100%	100%	SITI	72%	96%
HATI	100%	100%	HATI	100%	100%
KYUU	100%	100%	KU	68%	100%
Mean	94%	100%	Mean	85%	99%

られ、雑音の影響が取り除かれていることがわかる。

発音の仕方が互いに大きく異なった場合の認識を行うため、つぎの10単語

0(ZERO, REI), 1(ITI), 2(NI), 3(SAN),
4(YON, SI), 5(GO), 6(ROKU), 7(NANA, SITI),

8(HATI), 9(KYUU, KU)

を取り上げ、同様の認識実験を行った。

サンプリング周波数 4 kHz, S/N 比 35±5 dB の雑音環境下における認識結果を **Table 3** に示す。

発音の大きく異なる単語については音声のみによる

もの、口形情報を加えたもののどちらも 94%, 100% と高い認識率が得られている。一方、単語の中に (ITI, SITI), (SI, NI) など発音の似た単語がややあると音声のみのほうは認識率がすぐ低下してしまうが口形特徴も加えて認識を行うとその低下をおさえ高い認識率を保つことができる。

また発音の異なった単語でも、音声情報のみの認識では、同じ単語で発音の仕方が異なると、認識率が低下することもあるが、口唇の動きの情報を加えると、その低下を改善することができる。

4.4 検 討

雑音環境下では、音声情報が失われるほど単語認識率は悪くなる。最悪の場合でも口形情報のみで認識を行った場合、他の実験によって特定話者、限定 10 単語で認識率 90% 以上という結果を得ている⁸⁾。発音の仕方の類似した単語群の認識実験において、雑音の S/N 比と認識率の関係は Fig. 5 のようになった。音声特徴のみの認識では、S/N 比が 40 dB 以上の場合、ある程度のよい認識率が得られるが、それ以上に雑音が大きくなってくると、単語の認識率は急激に低下する。これに対し、口唇の動きの情報も加えた認識を行うと、その低下が改善される。このことから、雑音環境下では、口唇の動きの情報と音声情報を併用する、本論文で提案する認識法が有効であることがわかる。

また、発音の大きく異なる単語間の認識では、その認識率は、さらにより結果となる。

単語認識を行う場合、一般にその認識率は、サンプリング周波数が高いほどよくなると考えられる。しかし強い雑音環境下になると、音声特徴のみではサンプリング周波数を高くしてもそれほど認識率をあげることができなくなる。一方、口形の特徴を加えると視覚的に認識を補助することになりその認識率を高めることができる。

Fig. 5 Recognition rate vs. signal noise ratio

5. ま と め

本論文では、騒音の激しい環境のもとで、母音の類似した限定単語の認識を行うためのシステムを構成し、基礎的な単語認識実験を行い、その結果について報告した。雑音のある環境下で得られた音声データから LPC 分析により特徴パターンを作成し、さらに X-Y トラッカを用いて唇の一点の位置をリアルタイムに測定し、得られた座標データから口唇の動きの特徴パターンを作成する。そしてこれらの二つのパターンをもとに単語認識を行った。

本手法では、口形情報の特徴パターン作成において、X-Y トラッカが、マークされた口唇の一点の座標値をデジタル出力してくれるため、従来のような複雑な画像処理を行うことなく口形情報の特徴パターンを作成でき、従来の方法よりも簡単で高速化されている。

また、音声情報のみによる単語認識では、大きな雑音環境や、単語の母音が類似している場合、認識率がどうしても低下してしまうのに対し、本論文で提案する方法は、口形情報と音声情報を併用した認識を行っているため、これらの低下を大幅に改善することができる。また、口形の無意味な動きによる誤認も、音声情報併用により改善される。

本システムを用いた認識実験の結果、本論文で提案する認識方法は、カメラ装置を取り付けるなどの装置上の不便さは多少あるものの、音声情報のみによる単語認識での認識率の低下を、大幅に改善することができ、特に、雑音環境下において母音の類似した単語の認識に有効であることが確認された。さらにこの研究を基礎として、実環境に近い状況を設定し、X-Y トラッカの取付を工夫することによって実用化できるような研究を行うことも重要である。

参 考 文 献

- 1) A. Noda, K. Hatanaka, K. Kiriya, M. Fukui and T. Morita: Application of Voice Recognition Device for Data Collection in Noisy Work Site, JAPAN-U. S. A. Symposium on Flexible Automation, 779/786 (1986)
- 2) R. Vemula and E. Lee: A Microprocessor-Based Noise Canceller for the Cockpit, Proceedings of the IEEE 1986 National Aerospace and Electronics Conference, 4, 4-1399, 1323/1327 (1986)
- 3) 中臺, 菅村, 中津: 2 入力による雑音除去手法を用いた自動車内の音声認識, 電子情報通信学会音声技法, SP 89-81, 41/48 (1989)
- 4) Y. Fukuda and S. Hiki: Characteristics of the Mouth Shape in the Production of Japanese-Stroboscopic Observation, J. Acoust. Soc. Jpn. (E),

- 3-2, 75/91 (1982)
- 5) E. D. Petajan: Automatic Lipreading to Enhance Speech Recognition, IEEE Global Telecommunication Conference, GLOBECOM '84 Conference Record, 265/272 (1984)
- 6) P. C. Pandey, H. Kunov and S. M. Abel: A Speech Processor Providing Fricative and Low-Frequency Periodicity Information for Single Channel Cochlear Prosthesis, International Conference on ASSP 87, 3, 4-2425 1422/1425 (1987)
- 7) N. L. Hesselmann: Structural Analysis of Lip-Contours for Isolated Spoken Vowels Using Fourier Descriptors, Speech Communication, 2, 327/340 (1983)
- 8) 松岡, 古谷, 黒須: 画像処理による読唇の試み—母音口形の識別およびそれに基づく単語認識—, 計測自動制御学会論文集, 22-2, 191/198 (1986)
- 9) K. Kurosu, T. Furuya, K. Matsuoka and S. Takeuchi: Lip Reading by Fuzzy Logic, Preprints of Second IFSA Congress, 2, 470/473 (1987)
- 10) 呉, 田村, 光本, 河合, 黒須, 岡崎: 音声・口形特徴量を併用するニューラルネットを用いた母音認識, 電子情報通信学会論文誌, J73-D-II-8, 1309/1314 (1990)
- 11) B. P. Yuhas, M. H. Goldstein, Jr., T. J. Sejnowski and R. E. Jenkins: Neural Network Models of Sensory Integration for Improved Vowel Recognition, Proceedings of IEEE, 78, 1658/1668 (1990)
- 12) 田中, 迫江: 音声認識におけるパターンマッチング手法, 計測と制御, 22-1, 99/103 (1983)
- 13) 田村, 黒須, 古谷, 松岡, 園山: 口形の画像情報と音声による単語認識, 第30回自動制御連合講演会前刷, 511/512 (1987)
- 14) 河合, 田村, 黒須: 音声・口形併用による自動券販売機の開発—駅名の認識—, 第37回情報処理学会全国大会, 1970 (1988)
- 15) T. Furuya, T. Matsuoka and K. Kurosu: Machine Operation through Human Voices and Lip Movements for the Noisy Circumstances, The 7th European Annual Conference on Human Decision Making and Manual Control, Paris, October 18, 63/68 (1988)
- 16) 杉山, 鹿野: WLR 尺度による単語音声認識, 電子情報通信学会論文誌 (D), J66-D-4, 385/391 (1983)
- 17) 中川, 坂本: 音声認識・話者認識のためのFFTケプストラムおよびLPCケプストラムの検討, 電子情報通信学会論文誌 (A), J66-A-12, 1199/1207 (1983)
- 18) 白井良明: パターン理解, オーム社 (1987)
- 19) 古井貞昭: デジタル音声処理, 東海大学出版会 (1985)