昭和63年度厚生省心身障害研究
「家庭保健と小児の成長・発達に関する総合的研究」

ヒューマン・インタフェースへの音声対話時の引き込み現象の応用に関する研究：うなずき反応を視覚的に模擬する音声反応システムの開発

（分担研究：相互作用と乳幼児の心理行動発達に関する基礎的研究）

渡辺富夫*、夏井武雄*

要約 本研究では、語りかけに対して聞き手がうなずくように適切なタイミングで反応する実用的な音声反応システムの開発を目指している。まず話し手と聞き手とが意識的に同期をとった場合について、話し手の音声と聞き手のうなずき反応との引き込み現象を分析している。次にこの分析結果に基づいて、入力音声に対するうなずき反応を推定する、マクロ層とミクロ層からなる階層型の音声-うなずき反応モデルを提案し、各種シミュレーション実験により、モデルの有効性を示している。最後に本モデルを組み込んだ、うなずきを視覚的に模擬する音声反応システムを、ＤＳＰを用いてパーソナルコンピュータで実現している。

見出し語：ヒューマン・インタフェース、引き込み現象、音声対話、音声反応モデル

1　緒言

生体リズムが相互に同期化する現象を引き込み現象という。人間同士の対話においても、語りかけに対して聞き手がうなずいて応答し、またそのうなずき反応に合わせて話し手が語りかけるといったように、話し手と聞き手とが相互に同期する引き込み現象が存在し、円滑な情報交換に重要な役割を果たしている。この話し言葉に動作が同期する現象は、出生後まもない新生児期にも観察され[1),2)]、人間にとって本質的な情報交換形態であると考えられる。したがって、この引き込み現象のメカニズムが人間-機械系に導入されるならば、人間と情報機械の円滑な情報交換が図られ、人間性を尊重したヒューマン・インタフェースの実現に役立つと期待される。

＊山形大学工学部
(Faculty of Engineering, Yamagata Univ.)

著者の研究室ではこれまでに、話し手の音声と聞き手のうなずき反応との同期現象における音声と動きの時間遅れの範囲を明らかにした[3)]。また、大量のデータ分析が可能な音声-体動同期現象の収録システム及び音声と体動の実時間分析システムを開発した[4)]。さらに、入力音声に対して適切なタイミングで発光ダイオードが点滅する音声反応システムのプロトタイプを開発し、機械側に適切なる反応をもたせることの有効性を確認した[5),6)]。西らも電話会話における相づちの発言促進効果に着目した研究を進めている[7)]。

本研究では、語りかけに対して聞き手がうなずくように適切なタイミングで反応する実用的な音声反応システムの開発を目指している。まず話し手と聞き手とが意識的に同期をとった場合について、話し手の音声と聞き手のうなずき反応との引き込み現象を分析している。ここでは、これまでの分析方法[1)]と異なり、うなずき反応も音声信号も２値信号として処理している。

次にこの分析結果に基づいて、うなずき反応を２層の音声時系列の線形結合で推定する階層型の音声 - うなずき反応モデルを新たに構築している。 最後にＤＳＰ(Digital Signal Processor)を用いて実用的なシステムを開発している。

2　音声対話時のうなずき反応実験

2.1　うなずき反応実験

意識的に話し手に同期させてうなずいた場合には、うなずき反応の個人差が小さいと考えられる。本実験ではあらかじめ話し手と聞き手とを設定して、 同時に二人の聞き手（L1, L2）が意識的に同期させてうなずいた場合について行なった。話の内容としては、留守番電話を想定しての講演依頼のメッセージを選定した。１回のメッセージの長さは約45秒であり、ほぼ同じ内容で２回行なった。さらに、L1とL2を交替した場合についても同様に実験を行った。うなずく動作は音声とビデオタイマの時間情報と共にビデオテープに収録される。

2.2　測定

うなずき反応については、うなずきの大きさよりもうずいているか否かに意味があると考えた。したがって、うなずき反応については、画像のフレーム（1/30秒間）周期単位でうなずき反応の有無により、うなずく動作期間をON区間(1)とし、 それ以外をOFF区間(0)としてうなずきのパラメータＭを定義した。一方、音声については 有声(ON)か無声(OFF)かのON-OFFパターンの構造に着目した。 音声信号のON区間とOFF区間の機械識別は、6/30秒の平均雑音レベルに12dB加えた値を臨界値として、画像のフレームと同じ判定周期（1/30秒間）についてこの値を越えた部分をON、越えない部分を OFFとして、音声のパラメータＶを定義した。

2.3　分析

L1, L2のＭとＶの時間的変化の例を 図１(a)(b) に示す。聞き手L1とL2のうなずき反応が類似しているのがわかる。両者のうなずき反応のずれの平均値差は0.12秒であり、意識的に話し手に同期させてうなずいた場合には、個人差が小さいことがわかる。ＭとＶの関係は、次式の相互相関関数で分析した。

$$C(\tau)=\frac{\sum_{i=1}^{n-30\tau}\{V(i/30+\tau)-\mu_V\}\{M(i/30)-\mu_M\}}{\sqrt{\sum_{i=1}^{n}\{V(i/30)-\mu_V\}^2}\sqrt{\sum_{i=1}^{n}\{M(i/30)-\mu_M\}^2}}$$

$$\tau=0,\ \pm1/30,\ \pm2/30,\cdots$$

このＭとＶについてτ(time lag)が－３秒から３秒まで 1/30秒毎にＣ(τ)を算出した結果 （クロスコレログラム）を 図１(c) に示す。τ＜0 の領域では、うなずく動作に対して音声が先行する領域であり、一方、τ＞0 の領域は音声に対してうなずく動作が先行する領域である。-1.2秒付近に高い山があり、聞き手が話者の音声に対して1.2秒後にうなずいている。 一方、0.3秒に深い谷がある。 これは、聞き手が休止（話しの区切り）を確認してからうなずき反応を開始しているのではなく、休止を予測してうなずいているのを示している。

自己相関関数からＭとＶは共に３秒の周期性を持ち、対話者相互のリズムが引き込まれている。うなずき反応を推定するには、この話（音声）のリズムを考慮する必要がある。ここでの３秒を周期とするリズムは、主として呼気段落区分での有声区間と休止区間とのON-OFFパターンにより構成される。無声子音の前に発生する無音区間は発声器官のメカニズム上生じる区間で、呼気段落区分では有声区間のON区間とみなすほうが適切である。したがって、2/30秒のフィルイン （2/30秒より小さな継続時間のOFF区間のみをON区間に置換すること）を施し、呼気段落区分でのON-OFFパターンを求める。また、本研究において１回のうなずき反応時間の長短は意味がなく、うなずき反応の有無に重要な意味を持つ。したがって１回のうなずき反応時間を１秒間に定めた。これは測定された平均的うなずき反応時間である。

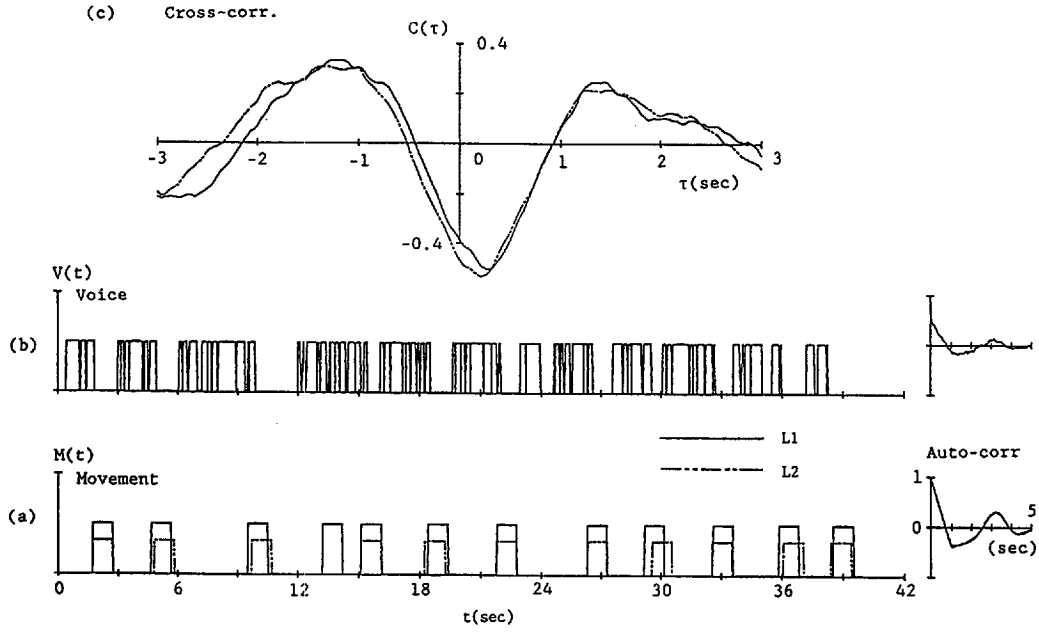

図1　MとVとの時間的変化 (a)(b) および、クロスコレログラム (c)

3　モデル化

3.1　音声-うなずき反応モデル

2章での分析結果より、うなずき反応は話の休止開始以前に開始していることが多いことが判明した。したがって、音声入力に対するうなずき反応を推定するモデル（音声-うなずき反応モデル）は、うなずき反応を休止開始以前に推定が可能なモデルでなければならない。本研究では過去の入力音声情報に基づいてうなずき反応を推定する、音声-うなずき反応モデルを構築した（図2）。

本モデルは、呼気段落区分でのON-OFF（有声区間と休止区間）パターンに基づいて呼気段落区間内にうなずき反応が存在するか否かを判定するマクロ層と、マクロ層での判定結果を受けて1/30秒単位の2値信号に基づいてうなずき反応の開始点を推定するミクロ層からなる2層の階層モデルである。即ち、まずマクロ層によりうなずき反応が存在する区間を判定し、その区間についてミクロ層によりうなずき反応の開始点を推定し、うなずき反応を出力する。

（1）　マクロ層

マクロ層は呼気段落区分での音声のON-OFFパターンのリズムに着目し、10秒前後と比較的長い音声時系列情報に基づいて、うなずき反応が“ある時間区間内”に存在するか否かを判定する層である。

うなずき反応が音声の休止開始前から始まっていることを考慮して、音声時系列を呼気段落区分でのON区間（有声区間）と その後のOFF区間（休止区間）からなるユニットに分割する（図3）。ユニット単位でのON区間が占める割合をユニット時間率と呼ぶことにし、このユニット単位ごとにうなずき反応が存在するか否かを判定する。第iユニットのユニット時間率R(i)を次式で定義する。

$$R(i) = \frac{T(i)}{T(i) + S(i)} \qquad (2)$$

T(i):第iユニットでのON区間長
S(i):第iユニットでのOFF区間長

図4(a)(b)にR(i)の系列と、その対応するユニットでのうなずき反応の開始点の有無を示すMu(i)（開始点が有れば1，無ければ0）の系列を示す。マクロ層でのMuの推定$\hat{M}_u$は、第iユニットのMu(i)を(i-1)ユニット以前のRの線形結合により推定するMA（Moving-Average）モデルであり、次式で表される。

$$\hat{M}_u(i) = \sum_j a(j) \cdot R(i-j) + u(i) \qquad (3)$$

u(i):noise

推定結果$\hat{M}_u$を図4(c)に示す。　マクロ層は、$\hat{M}_u(i)$ が あるいき値を越えた場合（いき値処理と呼ぶ）に 第iユニットにうなずき反応が存在すると判定し、次のミクロ層に処理を移す。

（2） ミクロ層

ミクロ層は、マクロ層によりうなずき反応が存在すると判定された呼気段落区分でのユニット単位について、より詳細で比較的短時間な音声時系列情報に基づいてうなずき反応の開始点を推定する。このミクロ層でのMの推定$\hat{M}_m$は、1/30秒周期でうなずき反応を音声の2値時系列信号の線形結合で推定するMAモデルであり、次式で表される。

$$\hat{M}_m(i) = \sum_j b(j) \cdot V(i-j) + u(i) \qquad (4)$$

u(i):noise

$\hat{M}_m(i)$ があるいき値を越えた場合、その時点をうなずき反応の開始点と判定し、最終的なうなずき反応$\hat{M}$を出力する。図5に$\hat{M}_m$の例を示す。パラメータbの個数はAIC（赤池の情報量基準）を参考に60個に決定した。これは過

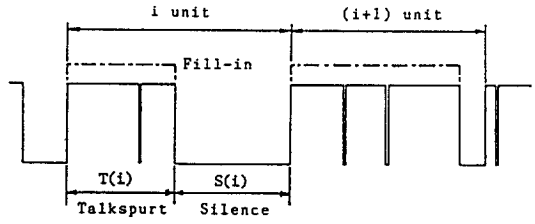

図3　呼気段落区分でのユニット時間率分割

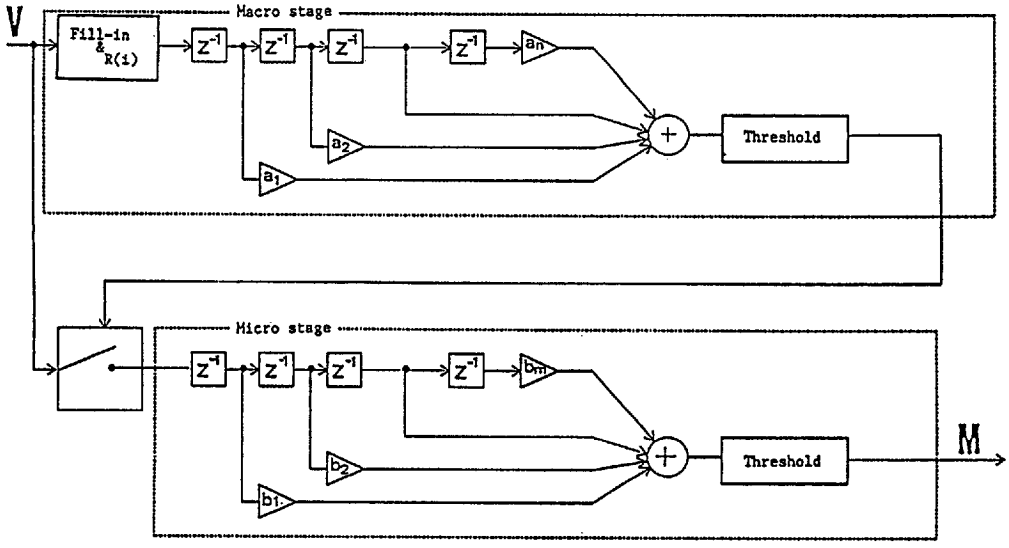

図2　音声-うなずき反応モデル（2階層モデル）の構成図

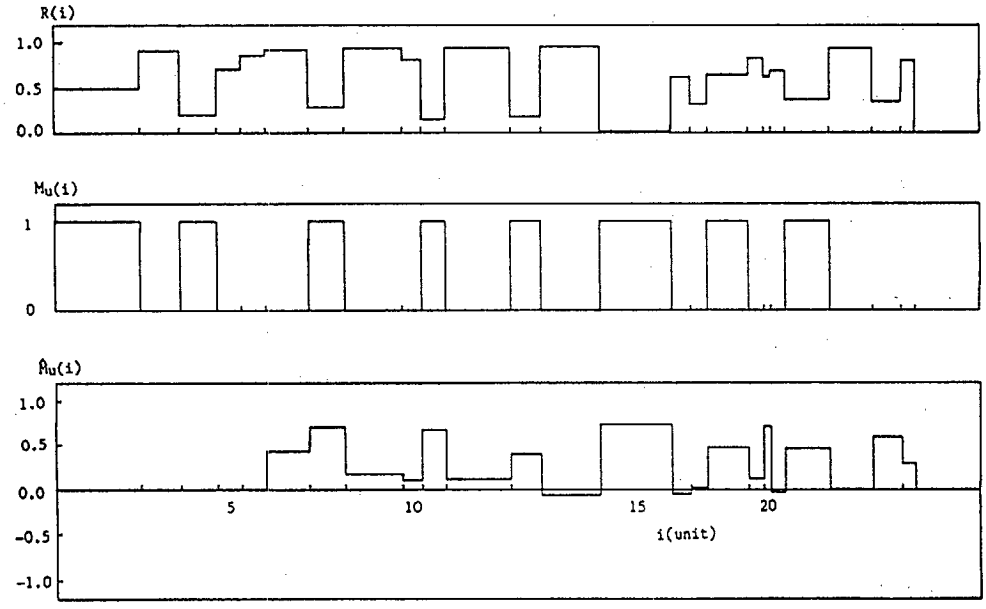

図4　R（i）、Mu（i）、$\hat{M}_u$（i）の時系列

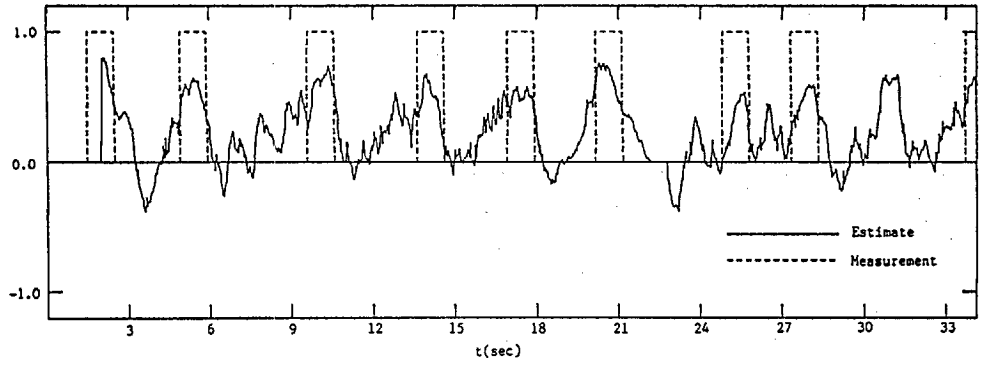

図5　ミクロ層におけるＭＡモデルでの推定結果

去２秒の音声時系列情報に基づいて推定することを意味する。ミクロ層での音声の２値信号に対してはフィルインを施していない。これは、フィルインを施した場合には、そのフィルイン時間だけ OFF区間の継続時間を計測する必要があり、実時間推定ができないからである。マクロ層ではユニット区間がONから始まるように設定してあるので、この問題はない。

3.2　モデルの評価

モデルの評価としてMと$\hat{M}$との時系列パターンの一致度を示す指標として相互相関 γ を用いる。ミクロ層でのいき値の変化に伴う相関による評価の結果を図６に示す。２層モデルにおけるマクロ層のいき値は最良の結果を与えるように設定してある。実線で示されたミクロ層の

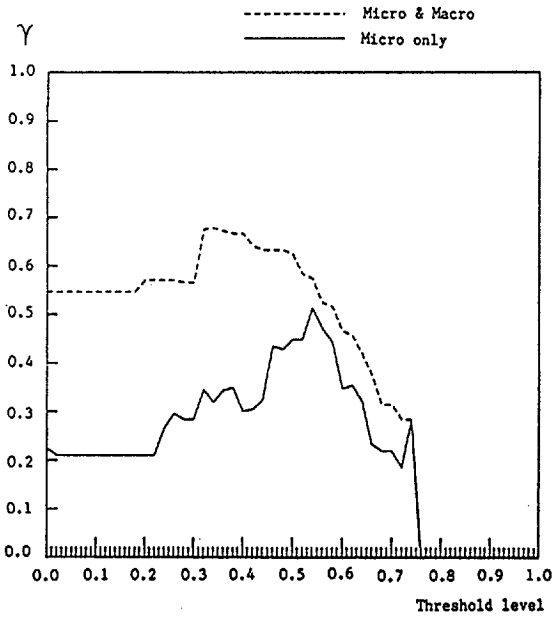

図6　ミクロ層でのいき値に対する相関γ

みによる推定結果については、ミクロ層でのいき値 0.34 のときγ = 0.32 なのに対し、破線で示された2層モデルによる推定結果ではγ = 0.68 と高い相関を示し、マクロ層が有効に作用している。このときの2層モデルによる推定結果を図7に、ミクロ層のみでの推定結果を図8に示す。マクロ層での推定は、ミクロ層での余分なうなずき反応の出力を抑えているのがわかる。なお、各うなずき反応ごとに、推定精度をある非線形関数を定めて評価した場合についても、2層モデルの有効性が確認された。

4　音声反応システム

前述の2階層モデルを組み込んだ音声反応システムを図9に示す。本システムは入力音声に対して適切なタイミングでレベルメータが点滅するシステムで、A／D変換ボード、DSPボ

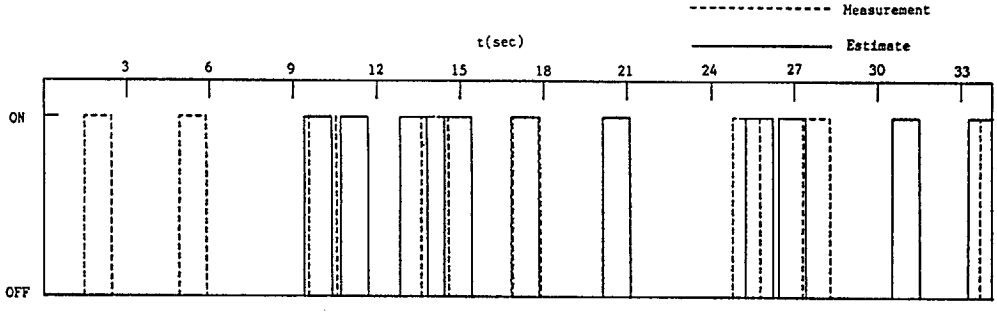

図8　ミクロ層のみによる推定結果

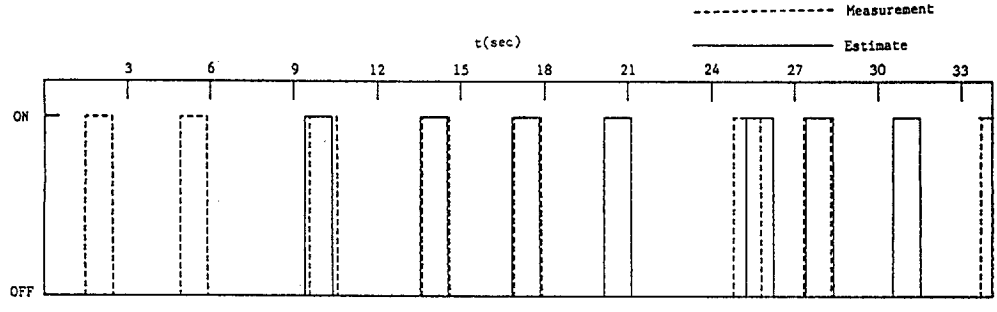

図7　2階層モデルによる推定結果

ードを搭載した パーソナルコンピュータ(PC9801VX)上で実現されている。A／D変換の制御、音声信号の２値化処理、ディスプレイへの出力は CPU 80286 が行い、 演算の高速性が要求されるユニット時間率の計算、及び２層モデルでの積和計算はＤＳＰ(TMS32010)において行われる。２つのＣＰＵは、共有メモリを通じてコミュニケートを行っており、実時間処理を実現している。

5　結言

本論文では、聞き手が意識的に話し手と同期をとった場合について、話し手の音声と聞き手のうなずき反応との引き込み現象を分析した。さらに、入力音声に対するうなずき反応を推定する、マクロ層とミクロ層からなる階層型の音声-うなずき反応モデルを提案し、 各種シミュレーション実験により、モデルの有効性を示した。最後に本モデルを組み込んだ、うなずきを視覚的に模擬する音声反応システムを、ＤＳＰを用いてパーソナルコンピュータで実現した。

参考文献

1) 渡辺,石井,小林:コミュニケーションにおけるエントレインメントのコンピュータ自動分析法, 医用電子と生体工学, 22-6, 419/425 (1984).

2) Kato, T. et al.:A Computer Analysis of Infant Movements Synchronized with Adult Speech, Pediat.Res. 17-8,625/628 (1983).

3) 渡辺富夫:成人間コミュニケーションにおけるエントレインメントの分析、情報処理学会論文誌,26-2, 272/277 (1985).

4) 渡辺富夫:対話における話し手の音声と聞き手のうなずく動作との同期現象の分析法、機械学会論文集、 C50-457,1128/1131 (1987).

5) 渡辺富夫:音声-体動同期現象のマン・マシン・インタフェースへの応用,情報処理学会論文誌, 25-2, 241/259 (1984).

6) 渡辺・結城:音声対話時のうなずき反応の分析とモデル化、第４回ﾋｭｰﾏﾝ･ｲﾝﾀﾌｪｰｽ･ｼﾝﾎﾟｼﾞｳﾑ論文集、157/162 (1988).

7) 西・小島:電話会話における「相づち」の発言促進効果、第３回ﾋｭｰﾏﾝ･ｲﾝﾀﾌｪｰｽ･ｼﾝﾎﾟｼﾞｳﾑ論文集、315/318 (1987).

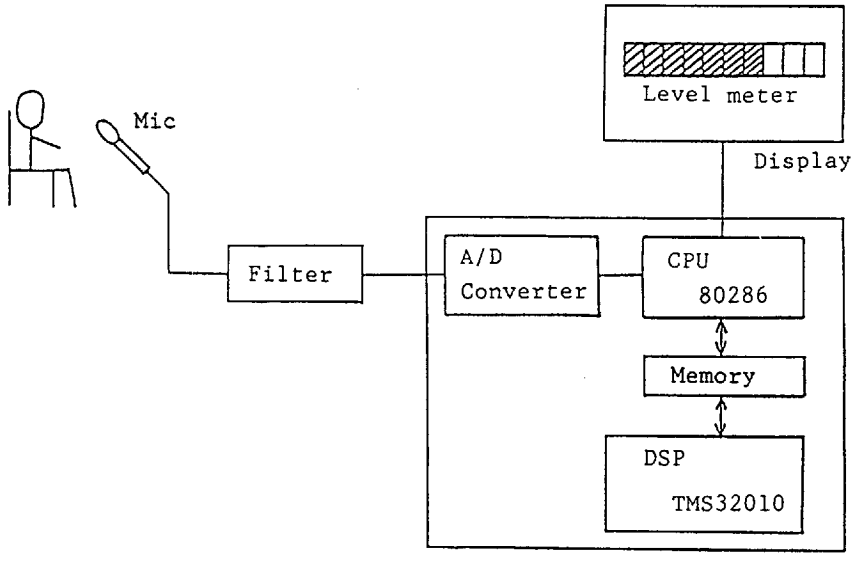

図９　うなずき反応システム概要図

ABSTRACT

A Voice Reaction System with a Visualized Response Equivalent to Nodding

Tomio Watanabe* and Takeo Natsui*

In human communication, a listener's movements such as nodding and facial expression are interactively synchronized with a speaker's words. This synchrony plays an important role in information exchange. It has been reported that this phenomenon is observed in a neonate in response to the mother's speech. The movement-voice synchrony is, therefore, an essential form of communication, and applications of this synchrony mechanism to a human-machine interface could be important for the smooth exchange of information in human-machine communication.

This paper first analyzes the synchrony between a speaker's voice and a listener's intentionally nodding in interpersonal communication. The voice and nodding are recorded on videotape while super-imposing a video timer with a frame unit of 1/30 second. The nodding parameter is defined in binary code according to whether or not a nodding movement is observed in each frame length of 1/30 second, while the voice parameter is also defined in binary code according to whether the speech power level is over or under a threshold value within the frame length. By obtaining the cross-correlation function between these two parameters, the authors reveal significant synchronous as well as lagged relationships between the speaker's voice and the listener's nodding.

Secondly, from the analysis of the voice-nodding synchrony, a hierarchical model of the voice reaction system is proposed. The model with simulated nodding consists of two stages -macro and micro, which respectively estimate the nodding response by the linear combination of the vocal binary code. The macro stage estimates the existence of nodding in each talkspurt plus silence duration unit with a fill-in operation of 2/30 second (2 frames) for the vocal binary code. This estimator is a moving-average model (MA model), which is expressed as the linear combination of units (the ratio of talkspurt duration to talkspurt plus silence duration) over approximately 10 seconds. Within the nodding episode estimated at the macro stage, the micro stage estimates the starting point of nodding in each frame length of 1/30 second (1 frame). This estimator is also an MA model, and is expressed by the linear combination of the vocal binary code over 2 seconds (60 frames). From the simulation for estimating the nodding response to vocal input, the effectiveness of this voice-nodding reaction model is demonstrated.

Finally, as an example of practical applications, the model is applied to a human-machine interface, and a new voice-nodding reaction system is developed. This system turns the light of a level meter on a graphic display on and off appropriately under the control of the optimum parameters of the model for vocal input. This study has many potential applications, in particular the realization of smooth information exchange in a human-machine interface.

検索用テキスト OCR(光学的文字認識)ソフト使用
論文の一部ですが、認識率の関係で誤字が含まれる場合があります

要約　本研究では、語りかけに対して聞き手がうなずくように適切なタイミングで反応する実用的な音声反応システムの開発を目指している。まず話し手と聞き手とが意識的に同期をとった場合について、話し手の音声と聞き手のうなずき反応との引き込み現象を分析している。次にこの分析結果に基づいて、入力音声に対するうなずき反応を推定する、マクロ層とミクロ層からなる階層型の音声一うなずき反応モデルを提案し、各種シミュレーション実験により、モデルの有効性を示している。最後に本モデルを組み込んだ、うなずきを視覚的に模擬する音声反応システムを、DSP を用いてパーソナルコンピュータで実現している。