人間-機械間の音声対話のための時間率適応化システム[†]

渡 辺 富 夫*

A Speech Activity Adaptation System to Improve Human-Machine Communication

Tomio WATANABE*

A machine which adapts to the speech of any human speaker and responds at an optimal utterance speed would be useful for the smooth exchange of information in human-machine communication, and the present study is concerned with the development of such a system. The author has already proposed a method for adapting machine conversational speed to speaker utterance speed. Based on sensory evaluation, this paper demonstrates that the speech activity adaptation itself is more essential for the smooth exchange of information than the adaptation to speech speed. Also, by analyzing the characteristics of speech activity estimation error as a function of observation time (talkspurt-silence period), the minimal number of talkspurt-silence periods required to estimate speech activity is determined. Finally, a conversational messaging system has been developed, which plays back stored speech segments controlled by the speech activity adaptation algorithm.

Key Words: human engineering, human interface, speech activity, human-machine communication, utterance speed

1. 緒 言

音声は、人間に適した、自然で容易な情報伝達手段であり、ヒューマン・インタフェースにおける有力な入出力である。人間機械間の音声対話においては、人間は機械へ音声入力する話し手であると同時に機械からの音声応答の聞き手である。人間にとって理解しやすい音声応答速度は、個人の発話速度（スピーチ区間のモーラ（注1）/秒）と密接な関係があることが知ら

れている¹⁾。したがって、個人の発話速度に対応して機械がその個人に最適な音声出力速度で応答できれば、個人と機械との円滑なコミュニケーションが図られ、人間に適合したヒューマン・インタフェースの実現に役立つと期待されよう。

人間機械間の音声対話における音声のON区間（有音区間：talkspurt duration）とOFF区間（無音区間：silence duration）の性質についての従来の研究としては、電話会話での中継回線の容量を等価的に増加せしめることを目的とした、音声ON-OFFパターンの統計的分析^{2)~4)}、連続音声の中の各種の区分の持続時間の性質^{5),6)}、音声波形における確率的発声終了検出法⁷⁾、会話音声の時間的特徴量に着目した遅延品質評価法⁸⁾、機械の音声出力タイミング⁹⁾などの研究がある。しかしながら、個人の発話速度に対応してその個人に適合した音声出力速度で応答する手法の基礎となるような研究は見当たらない。

時間率（speech activity）は、スピーチ区間内でON区間の占める割合をさす間接測定量であり、平均ON区間／（平均ON区間＋平均OFF区間）によって表現される。人を定めて発話速度を早口、普通、遅口の3種に変えた場合、ON区間の伸縮に比べてOFF区間の伸縮の比率が大きく、早口ほど時間率が高くなることが報告されている⁵⁾。しかしながら、早口にもかかわらずOFF区間を長く取る人もいるなど、スピーチのON-OFF特性には個人差があるので、ON-OFFパラメータから個人の発話速度を推定するには、人間一般についての発話速度とスピーチのON-OFFパラメータとの関係を明らかにする必要がある。したがって著者は前報において^{10)~12)}、人間一般についての発話速度とスピーチのON-OFFパラメータとの関係について検討し、時間率が発話速度と最も相関が高

[†] 第5回ヒューマン・インタフェース・シンポジウムで発表（1989・10）

* 山形大学工学部 米沢市城南 4-3-16

* Faculty of Engineering, Yamagata University, Yonezawa

(Received December 12, 1989)

(Revised May 16, 1990)

(注1) 1モーラは連続音声を変換した場合の1文字に相当する。スピーチ区間は一連の話のON区間からON-OFFを繰り返してON区間で終了するまでの区間である。

いことを明らかにした。さらに個人の発話速度を時間率に基づいて推定し、機械からの応答音声の OFF 区間を伸縮してその音声入力された時間率に一致させることにより、音声応答速度を適応化させる手法を提案した。

本論文では、まず官能検査により時間率適応化手法の有効性を評価している。つぎに応答音声を個人の時間率に一致させた場合と個人の発話速度に一致させた場合について官能検査により比較検討し、人間-機械間の音声対話において個人の時間率そのものへの適応化が発話速度への適応化よりも有効であることを明らかにしている。さらに入力音声の時間率を推定するのに必要な ON-OFF 区間回数について検討している。最後にこれらの知見に基づいて、音声対話のための時間率適応化システムのプロトタイプを開発している。

2. 官能検査による時間率適応化手法の評価

2.1 音声資料

発話実験としては、留守番電話を想定して、講演依頼の内容を被験者なりの言葉で伝達した発話実験 A と Table 1 のメッセージを定めた発話実験 B の 2 通りについて行った。発話実験 A では被験者ごとに言葉使いやスピーチの長さが異なり、19~23 歳の男性 46 人についてのメッセージのモーラ数の平均値±標準偏差は 129 ± 17.0 モーラであった¹¹⁾。一方、発話実験 B では 19~22 歳の男性 50 人全員が同一のメッセージを発するので、そのモーラ数は 44 モーラで一定である。

被験者の音声はまず VTR に PCM 録音される。その PCM 録音された音声はローパスフィルタを通してサンプリング周波数 10 kHz で A/D 変換され、パーソナルコンピュータ (PC 9801 VX 4) に収録される。その音声信号の ON 区間と OFF 区間の機械識別は、50 ms の平均雑音レベルに 12 dB を加えた値を臨界値として、ON-OFF の判定周期 10 ms 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 2 値化した。しかし無声子音の前の短い OFF 区間

は発声器官のメカニズム上生じる区間で、本目的には ON 区間とみなすほうが適切である。そのためにある値より短い継続時間の OFF 区間のみを ON 区間に置換する操作、いわゆるフィルインを施して、上記 OFF 区間を ON 区間に置換し、呼吸段落区分での ON 区間と OFF 区間を算出した。フィルイン時間は 130 ms に設定した。

2.2 時間率適応化手法

発話速度との相関はフィルイン 130 ms での時間率と最も高く、また音声応答速度の適応化は OFF 区間のみを制御することによって十分有効であることから、著者はつぎの時間率適応化手法を提案した¹¹⁾。この適応化手法は、ON 区間を変更せずに 130 ms 以上の OFF 区間のみを伸縮して個人の時間率に一致させる手法である。

いま、留守番電話のようにあらかじめ録音されている音声を入力音声の発話速度に適応化させる場合を考える。フィルイン 130 ms での応答音声と入力音声の時間率がおおの $\alpha_1 = T_1/(T_1+S_1)$ と $\alpha_2 = T_2/(T_2+S_2)$ のとき、両者の時間率が一致するように、

$$\begin{aligned} T_1/(T_1+S') &= T_2/(T_2+S_2) \\ S' &= T_1S_2/T_2 \end{aligned} \quad (1)$$

T_1 : 機械応答音声の平均 ON 区間

S_1 : 機械応答音声の平均 OFF 区間

T_2 : 入力音声の平均 ON 区間

S_2 : 入力音声の平均 OFF 区間

S' : 適応化後の機械応答音声の平均 OFF 区間

S_1 を T_1S_2/T_2 に伸縮することによって、すなわち、フィルイン 130 ms を施した (呼吸段落区分での) 各 OFF 区間を $(T_1S_2)/(T_2S_1)$ の比率で伸縮すれば適応化が図られる。

Fig. 1 に発話実験 A での 46 人の発話速度と時間率の関係を示す。図中の + 印は全被験者 46 人中、時間率が最大値、平均値、最小値の被験者 h, m, l の録音音声のおおのについて、時間率適応化手法を用いて本人以外の被験者 45 人の録音音声に適応化させた結果である。図中たとえば、h の録音音声を X の録音音声に適応化させれば、h の録音音声の時間率と発話速度は矢印で示される + 印になる (X の時間率に一致する)。

2.3 官能検査による評価

時間率適応化の有効性を官能検査により評価する。官能検査用の音声資料として Fig. 1 の h, m, l の録音音声を、またその判定者として h, m, l の被験者 [h], [m], [l] を選定した (判定者を [] で表わし、音声資料と区別する)。判定者 [i] に元の音声

Table 1 Selected message (originally written in Japanese).

"Hello, this is Suzuki speaking.
As to the starting time of the meeting
for the reports,
I'd like to set it at 3 p.m. tomorrow.
Is that OK?"

もしもし鈴木です。
論文の打ち合せですが
明日の 3 時にしたいとおもいます。
いかがですか。

Table 3 Results of sensory evaluation by groups in Fig. 2, comparing X_j with $X_j \rightarrow X_i$ by $[X_i]$

Group X					
Evaluator	$[X_i]$	$[X_h]$	$[X_{m1}]$	$[X_{m2}]$	$[X_{m3}]$
Speech					
X_i			$X_i \rightarrow X_{m1}$	X_i	$X_i \rightarrow X_{m3}$
X_h			$X_h \rightarrow X_{m1}$	$X_h \rightarrow X_{m2}$	$X_h \rightarrow X_{m3}$
X_{m1}	$X_{m1} \rightarrow X_i$	X_{m1}			
X_{m2}	$X_{m2} \rightarrow X_i$	$X_{m2} \rightarrow X_h$			
X_{m3}	$X_{m3} \rightarrow X_i$	$X_{m3} \rightarrow X_h$			

Group W		
Evaluator	$[W_i]$	$[W_h]$
Speech		
W_i		$W_i \rightarrow W_h$
W_h	$W_h \rightarrow W_i$	

Group Z		
Evaluator	$[Z_i]$	$[Z_h]$
Speech		
Z_i		—
Z_h	$Z_h \rightarrow Z_i$	

定して上記と同様な官能検査を行った。このグループの選定においては、グループ内で互いに発話速度および時間率の差が大きく、かつグループ内の音声資料のおのに対して発話速度あるいは時間率がほぼ等しい音声資料 ($q, X_{m1}, X_{m2}, X_{m3}, r$) が得られるように選定した。添字 h, m, i は、グループ内で時間率が高い、平均、低いを表す。官能検査結果を **Table 3** に示す。まずグループ X については、時間率が最小値の判定者 $[X_i]$ は、 X_{m1}, X_{m2}, X_{m3} の音声すべてに対して時間率適応化を好ましいと判定している。時間率が平均値の判定者 $[X_{m1}, X_{m2}, X_{m3}]$ は、 $[X_{m2}]$ の音声資料 X_i に対する判定以外は、時間率適応化を好ましいと判定している。時間率が大きい $[X_h]$ は、音声資料 X_{m1} に対する判定を除いて時間率適応化を好ましいと判定している。グループ W およびグループ Z についても時間率適応化を好む判定をしている（ただし $[Z_i]$ の判定は首尾一貫しなかった）。これらのことは、時間率適応化の有効性を示すもので、**Table 2** の結果とを総合して、時間率適応化は少なくとも入力音声の時間率が平均値以下の場合に有効であるといえる（ $[X_{m2}]$ の X_i に対する判定を例外とみなした）。

前論文^{10)~12)}では、音声応答速度の適応化を発話速度を指標にして検討した。しかし、時間率適応化それ自体が適応化の本質である可能性もある。すなわち、時間率適応化を行えば、必然的に応答速度が個人の発話速度に近づくと解釈するわけである。ここでは、音声応答速度の適応化の指標として時間率と発話速度とのどちらが適切であるかを、応答音声と個人の時間率に一致させた場合と個人の発話速度に一致させた場合

について官能検査により比較検討する。メッセージ文を定めた本実験では音声資料のモーラ数が一定なので、発話速度を一致させるには、音声資料 1 の 130 ms 以上の OFF 区間を伸縮して判定者の音声資料 2 の発話時間に一致させればよい。すなわち

$$S' = (S_1(N_m - 1) + H_2 - H_1) / (N_m - 1) \quad (2)$$

S_1 : 音声資料の平均 OFF 区間

H_1 : 音声資料の発話時間

H_2 : 判定者の音声資料の発話時間

N_m : フィルイン時間 m での ON 区間の回数 ($m = 130 \text{ ms}$)

S' : 適応化後の音声資料の平均 OFF 区間

フィルイン 130 ms での各 OFF 区間を S'/S_1 の比率で伸縮すればよい。

Fig. 3 に示したグループ ($X_i, X_{m1}, X_{m2}, X_{m3}, q$), ($Z_i, Z_h, X_{m1}, X_{m2}, X_{m3}$), (W_i, W_h, r) について、上記の発話速度適応化による音声と元の音声との一対比較および時間率適応化による音声と元の音声との一対比較による官能検査結果を **Table 4** に示す。たとえば、 q と X_{m1} とは時間率が近いので、判定者 $[q]$ に音声資料 X_{m1} と X_{m1} を q の発話速度に一致させた音声資料 $X_{m1} \rightarrow q$ を一対比較で聴かせた場合、 $[q]$ は X_{m1} を選定している。また q と X_i とは発話速度が近いので、 $[q]$ に音声資料 X_i と X_i を q の時間率に一致させた音声資料 $X_i \rightarrow q$ を一対比較で聴かせた場合、 $[q]$ は $X_i \rightarrow q$ を選定している。したがって時間率が平均値で、発話速度が遅い判定者 $[q]$ は、同じ発話速度よりも同じ時間率の音声を好む判定をしている。時間率も発話速度も平均値である判定者 $[X_{m1}], [X_{m2}]$,

Table 4 Results of sensory evaluation by groups in Fig. 2, comparing speech activity adaptation with speech speed adaptation

Group ($Z_1, Z_h, X_{m1}, X_{m2}, X_{m3}$)			
Speech \ Evalu-ator	$[X_{m1}]$	$[X_{m2}]$	$[X_{m3}]$
Z_1 $Z_1 \rightarrow X_1$	$Z_1 \rightarrow X_{m1}$	Z_1	$Z_1 \rightarrow X_{m3}$
Z_h $Z_h \rightarrow X_1$	Z_h	Z_h	Z_h
Group ($X_1, X_{m1}, X_{m2}, X_{m3}, q$)		Group (W_1, W_h, r)	
Speech \ Evalu-ator	$[q]$	Speech \ Evalu-ator	$[r]$
X_1 $X_1 \rightarrow q$	$X_1 \rightarrow q$	W_1 $W_1 \rightarrow r$	W_1
X_{m1} $X_{m1} \rightarrow q$	X_{m1}	W_h $W_h \rightarrow r$	$W_h \rightarrow r$
X_{m2} $X_{m2} \rightarrow q$	X_{m2}	→Speech activity adaptation →Speech speed adaptation	
X_{m3} $X_{m3} \rightarrow q$	X_{m3}		

$[X_{m3}]$ は, $[X_{m1}]$ の Z_1 に対する判定を除いて同じ時間率の音声を好む判定をしている. 時間率が高い判定者 $[r]$ は, 同じ発話速度を好む判定をしている. しかしこの判定は, 時間率適応化を図った音声資料が W_h の発話速度に相当するきわめて早い発話速度になったために生じた結果であり, $[r]$ は早すぎる発話速度を好まなかったと推定される. これらの結果より, 少なくとも時間率が平均値以下の場合には, 発話速度よりも時間率そのものに適応化させることのほうが好ましいことが明らかになった. したがって発話速度に適応化させるのではなく, 時間率そのものに適応化させることが円滑な対話を図る本質であるといえる.

Fig. 4 Machine's speech activity after adaptation for speaker's speech activity

システム開発においては, 機械にあらかじめ録音する音声を平均的な時間率 (0.64~0.66) になるように設定し, 入力音声の時間率がそれ以下ならば時間率適応化を図り, それ以上の場合には録音音声をそのまま出力するように設定すればよい (Fig. 4).

3. 時間率適応化システム

3.1 時間率推定

適応化システムにおいて, 入力音声の時間率をそのスピーチ区間の途中で推定できれば, システムは直ちに応答処理のための準備を進めることができる. ここでは, 入力音声の時間率を安定して推定するのに必要な ON-OFF 区間の回数について検討する.

Fig. 1 の発話実験 A での音声について, ON 区間に囲まれた OFF 区間の回数 i までの時間率 α_i の時系列的推移を Fig. 5 に示す. α_i は $i=8$ 以後では最終の時間率に近い値で安定している. 本論文では最終

Fig. 5 Speech activity (α_i) as a function of the number (i) of silence durations

の時間率として許容される範囲として最終時間率の ± 0.03 の範囲を選定した。これは同一被験者が Table 2 のメッセージを複数回繰り返したときの時間率の標準偏差が約 0.02 (ある被験者は 12 回で 0.018, 別の被験者は 8 で 0.022) であったので, その 1.5 倍を目安にした。Fig. 5 での 46 例の音声について, α_i がこの許容範囲に安定的に収まる OFF の回数 i を計測した結果を Table 5 に示す。46 例について $i=8$ では 76.1%, $i=9$ では 82.6% が許容範囲に収まる。

3.2 システム開発

上記の知見に基づいて, あらかじめ録音されている音声を個人の時間率に適応化させる, 音声対話のための時間率適応化システムのプロトタイプを開発した。録音音声は Table 6 に示すメッセージであり, その時間率は平均的な時間率 0.64 である。留守番電話を想定した実験での最大の OFF 区間は 2.5 秒であったので, これ以上の OFF 区間はスピーチの中断と判断し, それを除去して時間率を計算する。時間率適応化の際には, 録音音声の ON 区間開始前 30 ms の時点に OFF 区間を挿入し, 無声子音の欠落を除去する¹³⁾。時間率を計算する OFF 回数は 8 回に設定した。

システムは入力音声に対して 8 回の OFF 区間を含む ON 区間の終了時点で時間率を計算し, その時間率が録音音声の時間率 (0.64) 以下ならば時間率適応化を図り, 録音音声の時間率以上ならばそのままの録音音声を入力音声終了後に音声出力する。本システムを留守番電話を想定してメッセージを伝達する実験で官能検査した結果, 被験者 5 人全員が好ましいと判定し, システムの有効性を確認することができた。

Table 5 The number (i) of silence durations in which the i -th speech activity (α_i) ranges within ± 0.03 of the final speech activity. The i -th speech activity is calculated during the period from the onset to the i -th silence duration

	Mean	Standard deviation
OFF number (i)	5.7	2.6
Total OFF number (N_m-1)	13.6	3.9

Table 6 A stored message in the speech activity adaptation system (originally written in Japanese). "I received your message. Yes, I will inform Ishii. Thank you for calling."

はい わかりました。
確かにメッセージを
石井に伝えます。
電話ありがとうございました。

4. 結 言

本論文では, 人間機械間の音声対話の円滑化を目的として, 人間の音声の時間率に機械応答音声の OFF 区間を伸縮させて一致させる時間率適応化手法の有効性を官能検査により評価し, 本手法の適用時間率範囲を明らかにした。その際, 適応化の指標として発話速度よりも時間率そのものがより適切であることを示した。さらに, 音声対話のための時間率適応化システムのプロトタイプを開発し, その有効性を確認した。

音声情報の時間構造 (ON-OFF パターン) だけでなく, ピッチ構造, 振幅構造の韻律情報も考慮した音声対話のための適応化手法の確立, ならびに種々の言語にわたる時間率適応化手法のロバスト性の評価が今後の課題である。

最後に, 本研究は西條智幸君 (現在 NTT) の協力を受けたことを付記し謝意を表する。

参 考 文 献

- 1) 渡辺富夫: マン・マシン・インタフェースにおけるスピーチの理解しやすい速度に関する研究, 日本機械学会論文, 53-C-496, 2705/2710 (1987)
- 2) P. T. Brady: A Statistical Analysis of On-Off Patterns in 16 Conversations, Bell Syst. Tech. J., 47-1, 73/91 (1968)
- 3) J. G. Gruber: A Comparison of Measured and Calculated Speech Temporal Parameters Relevant to Speech Activity Detection, IEEE Trans. Communications, C-30-4, 728/738 (1982)
- 4) H. H. Lee and C. K. Un: A Study of On-Off Characteristics of Conversational Speech, IEEE Trans., C-34-6, 630/637 (1986)
- 5) 比企, ほか: 連続音声の音韻区分の持続時間の性質, 電気通信学会誌, 50-5, 69/76 (1967)
- 6) 比企静雄: 連続音声の各種の区分の持続時間の性質, 電子通信学会誌, 50-8, 99/104 (1967)
- 7) 西, ほか: 音声対話における確率的発声終了検出法, 電子情報通信学会誌, 70-D-11, 2108/2114 (1987)
- 8) 伊藤, 北脇: 会話音声の時間的特徴量に着目した遅延品質評価法, 音響学会誌, 43-11, 851/857 (1987)
- 9) 西, ほか: 人間機械間の音声対話におけるタイミング, 第 2 回ヒューマン・インタフェース・シンポジウム論文集, 291/294 (1986)
- 10) T. Watanabe: A Method for Adapting Machine Conversational Speed to Speaker Speed in Man-Machine Communication, 第 27 回 SICE 学術講演会予稿集, 914/922 (1988)
- 11) 渡辺富夫: マン・マシン・インタフェースにおける音声対話速度の適応化, 第 4 回ヒューマン・インタフェース・シンポジウム論文集, 303/310 (1988)
- 12) T. Watanabe: The Adaptation of Machine Conversational Speed to Speaker Utterance Speed in Human-Machine Communication, IEEE Trans. Systems, Man, and Cybernetics, SMC-20-2, 502/507 (1990)
- 13) J. Hirai, et al.: A Voice Message Store and Forward System Based on the DSP Technique, Trans. of the IEICE, 71-9, 876/880 (1988)