音声言語情報処理 22－13
（1998. 7. 25）

非同期型音声メッセージシステムの提案

幸 英浩　西本 卓也　新美 康永

京都工芸繊維大学 工芸学部 電子情報工学科

〒606-0962 京都市左京区松ヶ崎御所海道町

http://www-vox.dj.kit.ac.jp　yuki@vox.dj.kit.ac.jp, nishi@dj.kit.ac.jp

あらまし 非同期型音声メッセージシステムにおいて自然なコミュニケーションを行なうには、話し言葉の漸次性を利用できるシステムが必要になる。また、蓄積されたメッセージを有効に利用するためには、仮想会議の参加者や傍観者など、聞き手の立場に応じて異なる方法で再生される必要がある。また、音声認識等の技術を用いて会話音声の要約や書き起こしなどが行われることが望ましい。本稿では、非同期な音声対話を実現する方法として、音声メッセージの再生中に聞き手の応答音声を記録するシステムを提案し、このシステムにおける音声メッセージの蓄積方法および再現方法について検討した。

キーワード ボイスメール，非同期型メディア，AVM，音声認識

An Asynchronous Voice Messaging System

Hidehiro YUKI, Takuya NISHIMOTO and Yasuhisa NIIMI

Department of Electronics and Information Science, Kyoto Institute of Technology

Matsugasaki, Sakyo-ku, Kyoto, 606-0962 Japan

http://www-vox.dj.kit.ac.jp　yuki@vox.dj.kit.ac.jp, nishi@dj.kit.ac.jp

Abstract To communicate efficiently with asynchronous voice messaging systems, we must take into an account the ill-formness and incrementality of the spontaneous speech. The messages must be concatenated and played in many ways to meet the users' demands. It is also desired to make the summaries of the messages using both speech and text. In this paper we described an asynchronous voice messaging system, which records the listener's responses which overlap the original voice to make the conversation smooth. We also proposed some methods to playback the messages in the system.

Key words Voice mail, Asynchronous media, AVM, Speech recognition

1 はじめに

近年、手紙・電話・ファクスなど既存の通信手段を補完するメディアとして、電子メールが急速に普及しつつある。電子メールは時間と空間を越えたコミュニケーションが可能であるため、従来の社会組織にとらわれない新たなコミュニティを自由に構築することができる。

電子メールが広く使われている理由として挙げられるのは、相互互換性が高く、比較的短時間でメッセージが届く、といった点が挙げられるが、メディアとしての電子メールには次のような特長があると考えられる。

1. **非同期型メディア**であり、通信相手の時間を拘束しないので、お互いの時間を有効に使うことができる。また、時間をかけて返答を作成することができるため、複雑な内容のやりとりを行うことが容易である。また、同報が容易であり、記録を残しやすい。
2. **テキストメディア**であるため、キーボード操作を訓練することにより、比較的容易に入力が行える。また、伝達する内容に曖昧さがない。さらに、メッセージを閲覧したり検索することが容易である。また、他の文字メディアから情報を転用したり、印刷することで非電子メディアに変換できるなど、情報の加工や再利用が容易である。

一方で、電子メールの不便な点としては、感情を伝えにくいことが挙げられる。また、書き言葉で文章を作成することは、音声会話などと比較すると繁雑な行為である。したがって、些細な要件を文章にするのが億劫であったり、積極的に発言することに抵抗を持つこともある。このような理由から、現在の電子メールは必ずしも誰もが気軽に発言できるメディアではないと考えられる。

本稿では、これら電子メールの利点や欠点をふまえて、ユーザが発言しやすい音声メッセージシステムのあり方を検討する。そして、文字メディアと音声メディアを統合しつつこれらの要件を満たすシステムとして AVM[1] を提案し、格納されたメッセージを音声合成や音声認識などを用いて再現する方法について検討する。最後に、我々が現在構築中のシステムについて述べる。

[1] Asynchronous Virtual Meeting, Asynchronous Voice Messaging

2 話し言葉の漸次性

話し言葉の大きな特徴としては「漸次性」を挙げることができる。つまり、相手の知らない情報を段階的に伝達したり、自分が伝えたい内容の断片を思いついた順序で発することが可能である [1]。

また「共話」とよばれる現象も指摘されており [2]、対話者が共同で一つの話を作っていくものとして対話をとらえることができる。共話においては、相づちや連話などの発話行為は、相手の反応を確認し、それに基づいて自分のメッセージを詳細化する役割を持っている。

漸次的なコミュニケーションは、話者の心理的負担を軽減していると同時に、「いま誰かとコミュニケーションしている」という臨場感を醸し出している。これにより自然な会話が成り立っている。

一方で、留守番電話にメッセージを録音するような場合には、相手の反応を聞きながら喋ることはできない。そのためあらかじめ話したい内容を頭の中で文章化する必要性に迫られる。このような形式でのメッセージの伝達では音声コミュニケーションの利点は生かされていない。

また、対話的な自然発話においては、相手発話の終了前に喋り始める、いわゆるオーバーラップ発話が頻繁にみられる。RWC プロジェクトによって収録された 1 対話においては、応答発話中の過半数が相手発話にオーバーラップしており、全体の対話時間は、2 話者の発話時間の合計に対して 13% 短い時間であった [3]。

音声メッセージシステムにおいてユーザの漸次的発話を許容しオーバラップ発話を記録することにより、話し言葉を柔軟に受理することができる。これにより自然な会話を実現することが本研究の目的である。

3 既存の非同期型音声システム

ここでは、現在用いられている非同期・蓄積型の音声メッセージシステムについて述べる。

3.1 留守番電話

留守番電話は、非同期の音声メッセージシステムの一番簡単な例である。しかし、留守番電話にメッ

セージを伝える時には、発話に応答する「聞き手」が存在しない。また、留守番電話の応答には2章で述べた話し言葉の「漸次性」が存在しない。さらに、メッセージは録音された順に保存されるのみであり、メッセージ間の関係を意識しながら会話を継続することは困難である。

企業等においてはボイスメールシステムや、電子メールを電話操作で読み上げるComputer Telephony Integration (CTI) システムなども導入されはじめているが、これらにおいても同様の問題があると考えられる。

3.2 マルチメディア対応電子メール

電子メールソフトのバイナリファイル添付機能を用いて音声ファイルを送受信することができる。また、いくつかのソフトは音声添付ファイルの録音・再生を行うための専用インタフェースや、高能率の音声符号化機能を持っている[2]。

しかし、これらのメーラーは従来のテキストメール環境を元にして設計されているため、音声を気軽に入力できるインタフェースではない。また、音声ファイルの内容を確認するためには1ファイルづつ再生操作を行う必要がある。

3.3 音声認識による電子メール入力

近年、汎用の音声認識ソフトウェアや、音声認識機能を統合した電子メールソフトが製品化され、注目されている[3]。

しかし、入力した音声メッセージが持っていた声質や感情などの非言語情報は除去されてしまうため、せっかくの豊かなコミュニケーションの可能性を切り捨てている。また、漸次的に発話されたメッセージの多くは、文法に従っていなかったり、言い誤りや不要語を含んでいたり、倒置が多かったりする。このようなメッセージがテキスト化された場合、発話者にとってもメッセージの受け手にとっても不自然に感じられる。また、いわゆる「話し言葉」を音声認識システムがどの程度受理できるか、ユーザにとって明確ではない。

このようなことから、ユーザが話し言葉でメッセージを容易に作成するためには、新たなインタフェースに基づいたシステムが必要である。

[2] 例えばBecky! Internet Mailなど

[3] 例えばIBM ViaVoiceやNEC「しゃべっていいメール」など

4 メッセージの録音とデータ構造

我々は、音声の持つ自発的発言の容易性と、テキストが持つメッセージ可視性や検索しやすさを兼ね備えたシステムとして、AVMシステムを開発している。非同期型システムでありつつリアルタイム型システムと同様の臨場感を実現することが本システムの目標である。

4.1 新規メッセージの記録

メッセージ発言の場として会議室空間が定義され、発言は常に会議室空間に対して行われる。音声メッセージは始終端検出によって無音部分が取り除かれ、個々の音声区間に区切られて記録される。このときの音声区間を部分発話と呼ぶ。

このようなデータ形式を用いることで、録音時の無音データを削除することが可能となり、長時間のメッセージを少容量で送信することができる。

新たな会議室空間における最初の部分発話は、その会議室空間における時間軸の原点となり、それ以降の発言はすべて、この部分発話に対する相対的な時間関係に基づいて記録される。

1通のメールとして送信されるデータはテキスト，部分発話，付加情報の3種類の要素から構成される。

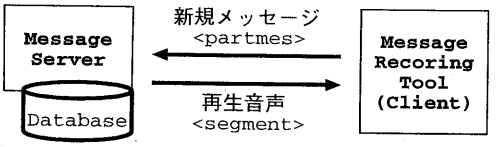
テキスト 既存の電子メールクライアントによって表示可能なテキストとする。

音声ファイル 1メッセージ内の複数の部分発話を1つの音声ファイルに結合してエンコードする。この情報は、音声添付ファイルに対応した既存の電子メールクライアントからは1通のメールごとに1つの音声ファイルとして再生できる。

付加情報 XMLに準拠した記述言語AVMLを定義し、各部分発話に関する情報を記述する。

AVMLは部分発話タグ (partmes) またはセグメントタグ (segment) の記述に用いられる。部分発話タグは新規および応答メッセージを登録するために使用される。また、セグメントタグは再生メッセージに付与され、新たな応答メッセージによって参照される情報を含む(図1)。

部分発話タグとセグメントタグはともに単語ラベルタグ (wordlabel) を含むことができる。


Message Server
Database
新規メッセージ
<partmes>
再生音声

Message Recoring Tool (Client)

図 1: AVML の利用方法

単語ラベルは後述する音声要約やメッセージ一覧表示などに用いられるもので、特定の音声区間に対応するテキストを表す。部分発話に対して連続音声認識やワードスポッティング処理を施すことで与えられるが、これらは送信時に付与される場合とメッセージを蓄積するサーバーで行われる場合が考えられる。

図 2 に新規音声メッセージの付加情報の例を示す。

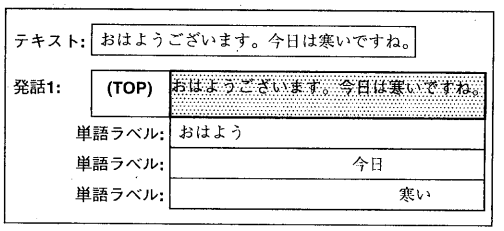


図 2: 新規音声メッセージの付加情報

4.2 応答メッセージの作成

蓄積型システムにおいて漸次的な発話行為を行なえるようにするため、メッセージの録音と再生を同時に行なえる全二重オーディオ機能を用いる。ユーザは音声メッセージを再生しながら、いつでもその音声に割り込んで相づちや返答を行なうことができ、これらは新たなメッセージとして保存される。

応答発話は元の発話にオーバーラップすることを前提にしている。既存のメッセージの再生時にも応答音声の始終端検出を行い、ユーザが発話した場合は再生中の既存メッセージと新たな発話とのタイミング関係を記録する。図 3 にユーザの応答例を示す。

応答メッセージを 1 通のメールとして送信する場合のデータ形式は新規メッセージの場合とほぼ同様である。ただし、各部分発話が直前のどの部分発話 (parent) のどの時刻 (reltime) に対する発話であるかを示すリンク情報が追加される。図 3 における話者 B の応答について、メッセージの付加情報を図 4 に、メールとして送信される情報を図 5 に示す。

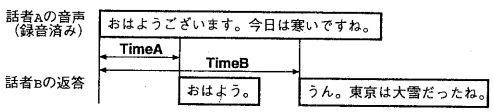


図 3: 音声メッセージに対する応答操作の例

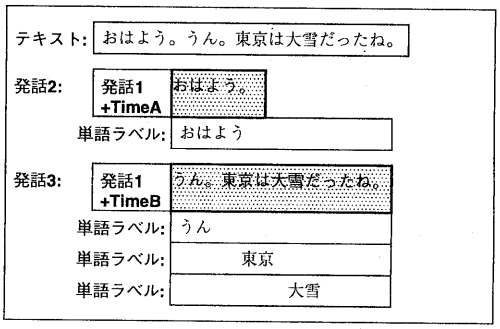


図 4: 応答メッセージの付加情報

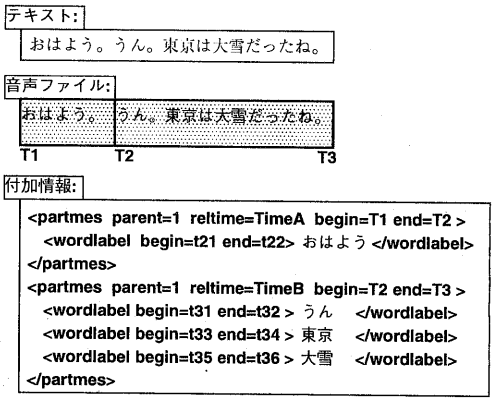


図 5: 応答メッセージの送信メール構造

1 つの会議室空間は、1 つの先頭部分発話と、これにリンクする複数の後続部分発話から構成される (図 6)。

5 メッセージの動的レイアウト

会議室空間における音声メッセージは 4 章で述べたリンク構造に基づいて蓄積される。このデータを必要に応じて合成することにより、非同期に蓄積さ

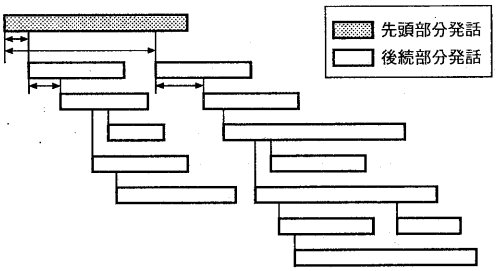


図 6: 部分発話のリンク構造の概念図．発話間の相対時間 (矢印) が記録される．

れた音声メッセージを仮想的な同期音声会話に変換する。

再現されたメッセージに新たな返答音声を付加させたい場合は、図 7 に示す情報を同時に生成する。これは再生中のデータの各時刻と、元のデータベース内のメッセージの対応づけをあらわしており、この情報を元にして新たな応答メッセージ (4.2 参照) を生成することができる。

```
<segment length=セグメントの長さ
         pmid=親メッセージID
         sender=親メッセージの送信者
         pmtime=部分発話内時刻
         playtime=再生開始からの時刻>
         begin=T1 end=T2 >
  <wordlabel begin=t21 end=t22 >
     おはよう
  </wordlabel>
</segment>
```

図 7: 会話音声の再生時に付与される情報

AVM は音声メッセージの記述方法、転送方法、リンク方法などを仕様として規定するが、音声データをどのように選択して再現するかは規定しない。AVM を構成する各要素を WWW に例えると表 1 のようになる。

表 1: WWW と AVM の比較

HTML	AVML
ブラウザ画面表示	AVM 再生音声
HTML エディタ	AVM 録音ツール

HTML 文書はブラウザの機能や表示オプションに応じて異なったレイアウトによって表示されるため、ユーザの環境や状況によってページの行幅や画像の有無等などが変化する。これと同様に、AVM システムにおいては、時間軸上におけるメッセージのレイアウトがユーザの環境や要求に応じて動的に変化する。

このような自由度を与えることにより、AVM は音声言語処理を応用するための共通のプラットフォームとなり得る。また、実用指向・エンターテインメント指向などさまざまな目的に応じたアイディアを実装することも考えられる。

音声メッセージの再生方法としては、次のような事例が挙げられる。

5.1 傍観者として会話を聞く場合

話者 C が話者 A、B の議論を第三者として聞く場合、非同期な音声を時間軸通りに仮想的な同期音声に変換して聞く方法がわかりやすい。

一方で、音声メッセージをテキストとして要約することにより、NetNews のように話者 A のメッセージ、話者 B の返答…といった一つ一つのメッセージを視覚的に取捨選択して、特定の部分のみを聞くこともできる。

5.2 メッセージの返答を聞く場合

話者 A が話者 B、C に対して「こんどの会議に参加できますか?」といったメッセージを送信し、会議の出欠を返答してもらいたい場合を考える。このような場合、話者 B、C は話者 A が送信したメッセージに対して、それぞれ非同期に自然会話で返答をつける。

話者 A が話者 B、C の音声を再生する場合、発話の時間軸通りに行なった再生では話者 B と C の音声が重なってしまい聞きづらいものとなる。

このような場合には、A, B, C の 3 人が同時に会話に参加しており、まず B が発言し、次に C が発言したかのように、再生タイミングをずらすことが必要となる。これにより、話者 B、C の音声が話者 A にとって聞き易い音声になる。

5.3 代理エージェントを用いた会話

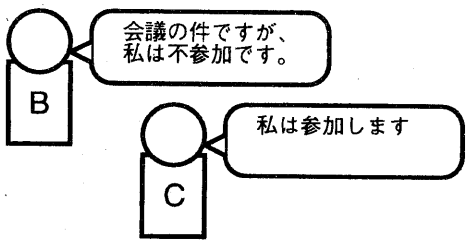
5.2 において、会議の出欠について A が聞くべき発言としては B:「会議の件ですが、私は不参加です。」C:「私は参加します」といった形式が自然である。

聞き手の立場を考慮しない再生方法を用いた場合は、A が会話を再生する場合に、自らの過去の発言も含めて聞くことになる。しかしこれでは「自分に対して話しかけている」という状況を実現することができない。

このような場合には、聞き手 A に対して B および C が必要なメッセージを伝えるための代理エージェントを用いることが考えられる。聞き手に応じてメッセージを取捨選択すると同時に、合成音声を用いて「会議の件ですが」という音声を B の応答音声に付加し、これを B および C の代理エージェントに語らせることによって、より適切に内容を伝えることができる (8)。

A: 会議の件ですが、どうですか?
B: あー、 私は不参加です。
C: うんうん、 私は参加します。

(1) 元の音声メッセージ



(2) 代理エージェントによる会話

図 8: 代理エージェントによる仮想会話の例

6 AVM システムの実装

現在、下記のようなサーバーとクライアントを実装している。

対話クライアント 全二重音声入出力が可能な Windows95/NT4.0 システムで動作し、基本的な電子メールの送受信機能と、音声を再生しながらオーバーラップ音声の録音を行う機能から構成される。音声再生中に聞き手が新たな発言を行うと、その発言中は再生音声が一時停止する機能を有する。

メッセージ受信サーバー UNIX システム上で動作し、E-mail として受信したメッセージの MIME デコードを行う。音声ファイルはケプストラム分析され、連続 DP 法によるワードスポッティングを行った上で、元ファイルと共にデータベースに格納される。

対話再生サーバー UNIX システム上で動作し、データベースに格納された音声から聞きたい話題をウェブの CGI を用いて選択する。

再生された音声にさらに返答をしたい場合は再生音声を対話クライアントに再度取り込んで返答操作を行う。

7 まとめ

音声通信と電子メールの統合は、電子メッセージシステムを有効に活用していく上で重要であると考えられる。本稿では、特に音声会話の漸次性を生かすことに配慮し、相槌などの音声オーバーラップを積極的に利用することで自然な音声会話を支援する非同期型メッセージシステム AVM を提案した。また、このようなシステムにおいて、同一の話題を電子掲示板的に閲覧したり、代理エージェントを用いて対話的に再現する方法についても検討を行った。

本稿で述べたデータ構造により、対話クライアントと対話再生サーバーの両方において拡張性の高い柔軟なシステムが構築できると思われる。また、音声に同期した映像を付加することで、非同期型のビデオ会議に拡張することも考えられる。

現時点では音声認識などの利用は予備的なものであるが、今後はシステムの実用性を高めるためにさまざまな改良を行っていく予定である。

参考文献

[1] 岡田美智男, 口ごもるコンピュータ, 共立出版, 1995.

[2] 伊藤昭, 矢野博之,「共話」－総発的対話の対話モデル, 情報処理学会研究報告 98-SLP-20-1, 1998.

[3] 西本卓也, 新美康永, 非同期音声メッセージシステムの設計, 計測自動制御学会：ヒューマンインタフェース部会誌,vol.13,No.1,pp47-54(1998)