会話の言い直しを利用したVoIPの音切れ補間方式の提案

小日向 肇[†,††] 星 徹[†] 松下 温[†]

インターネットを使用するVoIPにおいては，ネットワークでパケットの滞留が突然起こることによるスパイク性遅延変動により，500 msec以上の音切れが発生することがある．今まで提案されてきた音声補間方法は，短時間の音声欠損の補間を対象としていたため，補間しなければならない音韻が数個になる可能性が高い，スパイク性遅延変動による音切れを完全に補間することは困難であった．本研究では，通常の会話において“言い直し”がよく起こることであることに着目し，スパイク性遅延変動が発生したときに，受信側で言葉の一塊であるトークスパートを最初から言い直すことにより，音切れを補間する．この補間は，意味の伝達を確実にし，会話の自然さを保持することを特徴としている．このようなVoIPの音切れ補間方式である言い直し補間を提案し，さらに言い直し補間を実装したVoIP端末を開発し，評価を行い，有用性を確認した．

Voice Gap Compensation against Delay Spike Using Speech Repetition for Voice over IP

Hajime Obinata,[†,††] Tohru Hoshi[†] and Yutaka Matsushita[†]

Voice over Internet Protocol (VoIP) utilizes the Internet for voice communication. VoIP has a problem that a voice gap occasionally happens according to a delay spike. A voice gap caused by a delay spike is longer than 500 milliseconds. Voice gap compensation methods have been proposed before, but they are all only for short term gaps. No compensation method for long term voice gaps was proposed, because it was too difficult to compensate several phonemes. Our proposed method is for long term voice gaps. When a spike delay happens and next packet does not arrive, the receiver stops replaying voice and repeats the talk-spurt from the beginning instead of replaying next packet after the voice gap. This method can transmit the meaning of a sentence clearly, because automatic speech repetition simulates speech repair which often occurs in normal conversation. Humans are accustomed to this speech repair. This method can preserve naturalness as well, because it changes voice gaps into speech repetition. And the listener has little awareness of speech repetition compensation. We implemented the IP telephony system equipped with speech repetition compensation and also evaluated this method. As the result, we could ensure the usefulness of this method.

1. まえがき

音声通信にインターネットを利用するVoIPは，VoIP専用のネットワークを構築するなどのネットワークQoSの改善により，音切れや遅延などの問題を解決し，実用的なサービスが供されている．一方，経済性や自由度に勝る一般インターネット回線を利用したVoIPにおいても，その音質向上が切望されている．本研究では，ネットワークQoS制御を行わないベストエフォート型インターネット環境下において，主に受信側の対策によって，VoIPの音声品質向上を図ることを狙っている．

インターネットを使用するVoIPにおいては，ネットワークの状況により，パケットの伝送遅延が変動するため，受信側で音切れが起こらないような対策が必要となっている．このVoIPの音質に影響を与える遅延変動は，特性により2つに分けることができる．1つは，ほぼ一定の範囲内で，定常的にパケットの到着遅延が変動する現象である遅延ゆらぎであり，もう1つは，突然到着遅延が遅延ゆらぎより大きく増大し，その後すぐに到着遅延が徐々に元に戻りはじめる現象，言い換えると，しばらくパケットがまったく到着しなくなり，その後遅れていたパケットが次々に到着する現象であるスパイク性遅延変動である．このうち，遅延ゆらぎについては，ゆらぎ時間幅が比較的小さいた

† 東京工科大学コンピュータサイエンス学部
School of Computer Science, Tokyo University of Technology
†† 株式会社エイビット
Abit Corporation

め，受信側でプレイアウトバッファを設けて吸収しても，音質に影響の少ない範囲でパケットを廃棄することを許すと，遅延時間の増大は少ない[1]．一方，スパイク性遅延変動については，一般に発生頻度は比較的少ないが，遅延量は 500 msec 以上になることも多く，かつ発生を事前に予測することも困難である．これをプレイアウトバッファで完全に除去しようとすると，あらかじめバッファサイズを大きくして遅延を大きくとらなくてはならず，相互通話品質が低下してしまうという問題があった[2),3)]．またスパイク性遅延変動発生時に生じた音の途切れを，音声補間することが考えられるが，今まで提案されてきた音声補間方法は，数 10 msec 以内の短時間の音声欠損の補間を対象としていたため，補間しなければならない音韻が数個になる可能性が高い，数 100 msec 以上のスパイク性遅延変動による音の途切れを，完全に補間することは困難であった[4)~9)]．そのため，スパイク性遅延変動が発生すると，音の途切れや間延びになってしまい，意味の伝達が不完全となり，通話者に大きなストレスを与えていた．

本研究では，通常の会話において “言い直し”[10] がよく起こることに着目し，スパイク性遅延変動が発生したときに，スパイク直前に受信しているトークスパートを，最初から繰り返し再生することにより音切れを補間する．これにより，聞き手に対して，スパイク性遅延変動による音声の中断をカバーし，スムーズな音声の流れを確保する．受話者にとっては，“言い直し” が時々発生するが，話者が直接言い直したのと区別がつきにくく，会話の自然な流れを維持できることが期待できる．

また，この補間により発生する遅延の増大を，話速変換技術[11] により速やかに低遅延状態に戻すことにより，相互通話品質を落とすことなく補間を行うことを提案する．

本論文では，最初にスパイク性遅延変動の実測結果とその VoIP への影響について述べ，次に言い直し補間の基本原理と有用性について述べる．さらに，従来から提案されている短期補間と組み合わせて，条件により最適補間方法を選択していくハイブリッド補間について述べ，最後に評価結果について報告する．

2. スパイク性遅延変動とその VoIP への影響

一般インターネット回線では，スパイク性遅延変動が起こることが報告されている．これは，たとえば 60〜70 秒に 1 回程度などのある程度規則性が認められるものと，突発的に起こるものがあることが報告されている．また，スパイク性遅延変動が起こる前には，それを予測できるような前ぶれはまったく観測できないことと，スパイク性遅延変動発生時に，パケットロスがともなうことがあることが報告されている[2)]．

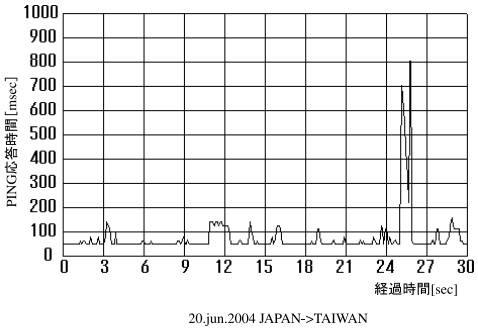


図 1 遅延測定結果
Fig. 1 Measurement result of a delay spike.

スパイク性遅延変動の原因としては，比較的規則的なものは，ルータのメンテナンスであることが報告されている．また突発的なものは，アクセスが集中したために，ルータでバッファリングされた場合や，アクセス集中や回線障害などで経路変更が起こるときに，それに時間を要するためと考えられている[2)]．

筆者らが PING を使用して，日本国内および世界各地のサーバとの間の応答特性を調査した結果でも，スパイク性遅延変動が起こることが確認できた．**図 1** に，PING を用いて応答特性を測定した例を示す．また応答特性を測定するときに，PING パケットにシーケンス番号を付与しておくことにより，パケットロスも検出しているが，スパイク性遅延変動が起こったときは，パケットロスもともなって起こる場合があることを確認した．また，スパイク性遅延変動が起こる前には，それを予測できるような前ぶれはまったく観測できないこともあわせて確認した．

したがって，従来スパイク性遅延変動による音切れを防ぐためには，常時プレイアウトバッファの遅延量を，スパイク性遅延変動量より大きくしておく以外に方法がなかった．しかしこの方法は，スパイク性遅延変動量が 500 msec 以上と大きいために，相互通話品質が著しく悪化してしまうという問題があった．3 章において，プレイアウトバッファを遅延ゆらぎを吸収するのに最適な遅延量に設定しているときに，スパイク性遅延変動が発生した場合の問題点を指摘し，それを解決する言い直し補間を提案する．

3. 言い直し補間方式の提案

3.1 スパイク性遅延変動による会話の途切れ

図 2 に例を示し，これを用いて提案方式を概説する．

図 2 (1) の会話で,「しんじゅく」の「しん」と「じゅく」の間でスパイク性遅延変動が発生した場合，遅れて到着したパケットを廃棄した場合は，図 2 (2) のように「じゅくえ」が欠損してしまうため，意味が分からなくなってしまう．また，遅れて到着したパケットを廃棄せず再生した場合でも，図 2 (3) のように「しん」と「じゅく」の間に無音が入るため,「しんじゅく」という 1 つの単語として理解できない可能性が高くなってしまう．このような場合は，受話者は「もう 1 度お願いします.」などと言って，送話者に再度同じことを言うように促すことになるため，会話の流れが一時的に中断され，無駄な時間が消費されると同時に，送話者と受話者双方にストレスを与えることになってしまう．また，意味が何とか伝達できたとしても，通常会話では発生しない無音区間ができるため，不自然な音声となり，受話者にはストレスになる．

3.2 言い直し補間方式による音声途切れの補間

提案する言い直し補間方式による再生では，図 2 (2) と (3) で起こった問題を解決するために，図 2 (4) に示すように,「しん」の後でスパイク性遅延変動により，次に再生すべき音声パケットが到着していない場合は，トークスパートの最初である「し」にもどって再生する．このようにする理由は，実際の会話では，次に話すべき言葉が見つからない場合や，痰がからんで言葉の発声が不完全となってしまった場合などでは，トークスパートの最初に戻って話す “言い直し” はよく起こることであるからである．したがって，人間はこの “言い直し” に慣れているため，“言い直し” によって補間しても，会話内容を理解するうえでは，スパイク性遅延変動が発生しなかった図 2 (1) と同等になるということが期待できる．また不自然な無音区間がなくなり，実際の会話でよく起こる “言い直し” になるため，受話者は相手が直接言い直したと認識することが期待できる．そのため，自然感が損なわれず，スパイク性遅延変動による補間に気がつきにくいということが期待できる．これらのため，図 2 (1) と (2) の場合のように，意味がうまく伝達されなかったために，受話者が相手に対して再度発声するように言葉で要求する必要がなり，したがって，会話がスムーズになり，また不自然な無音区間も解消できることが期待できる．

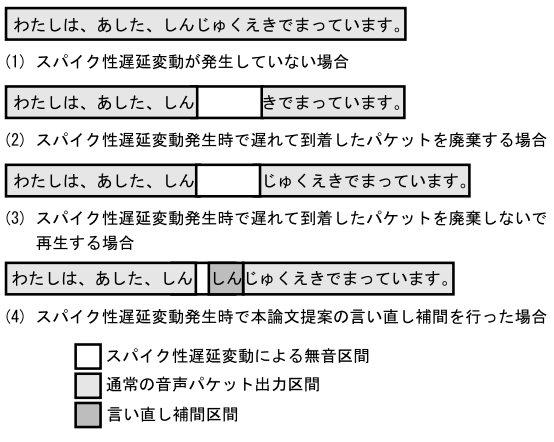


図 2 スパイク性遅延変動による音声遅れ発生時の処理方法による再生音声例
Fig. 2 Speech repetition compensation method.

3.3 提案方式の構成

図 3 に言い直し補間の基本構成を示す．通常の VoIP の受信機の構成に，トークスパート保持バッファ，有音判定部，トークスパート検出部を加えている．スパイク性遅延変動が発生していない通常の状態では，プレイアウトバッファに保持されている音声パケットを，音声デコーダに入力し，そのデコードされた音声を D/A コンバータに出力する．このときに，音声デコーダに入力すると同時に，このパケットをトークスパート保持バッファに保持する．またデコードされた音声は，有音判定部で，通常の周囲雑音下において，周囲雑音のみのとき（無音）と人間が発声したとき（有音）に，信号パワーに差があることを利用して有音と無音を判定する．具体的には，20 msec ごとにその区間の音声パワー（2 乗平均値）を求め，その値がスレッショールドを超えるときは有音区間と判定し，それ以下のときは無音区間と判定する．また上記 20 msec ごとに求めている音声パワーの最小値を記録しておき，この値を 5 倍した値をスレッショールド値とすることにより，周囲雑音の大きさに有音判定が影響されないようにしている．この音声パワーの最小値は，この値

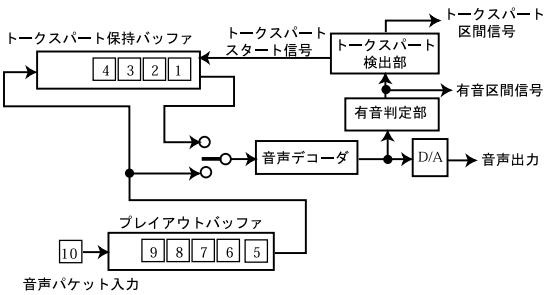


図 3 言い直し補間基本構成
Fig. 3 Speech repetition compensation architecture.

以下の音声パワーが観測されない場合は，徐々に大きくして，周囲雑音が増大しても最小値が追従できるようにしている．次に有音判定部の有音区間信号から，トークスパート検出部は，**図 4** に示すように，最小無音時間以上の無音区間の後の有音区間を，トークスパートのスタートとして判定する．この最小無音時間は，言い直し補間を行ったときに，意味的に違和感がないようにトークスパートが切断されるように，実験的に 100 msec とした．本報告において，トークスパート区間とは，最小無音時間以上の無音区間の後の有音区間の開始から，最小無音時間以上の無音区間が次に検出されるまでとする．ここで，有音区間とは，音声パワーがスレッショールド値を超える区間であり，無音区間はそれ以下の区間とする．また，トークスパート区間以外を非トークスパート区間とする．トークスパート検出部は，トークスパート区間を示すトークスパート区間信号を，トークスパートのスタートを示すトークスパートスタート信号とは別個に出力する．このトークスパート区間信号は，有音判定部の有音区間信号とともに，ハイブリッド補間と話速変換に使用される．次に，トークスパート検出部から出力されるトークスパートスタート信号によって，トークスパート保持バッファを消去する．これにより，トークスパート保持バッファには，新たに再生中のトークスパートの最初のパケットから順に保持される．スパイク性遅延変動が発生すると，音声パケットが到着しないため，プレイアウトバッファが空となってしまうため，トークスパート保持バッファよりパケットを取り出してデコードする．トークスパート保持バッファの音声をすべて出力した後に，スパイク性遅延変動により遅れて到着し，プレイアウトバッファに保持されていたパケットをデコードする．これにより，トークスパートの最初から連続して音声が再生されることになる．本方式はスパイク性遅延変動による 500 msec 以上の長期の補間に適用する．

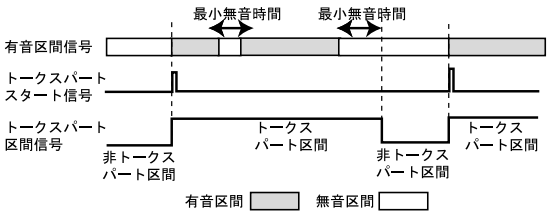


図 4 トークスパート検出部の動作
Fig. 4 The operation of talkspurt detector.

4. ハイブリッド補間

本章では，さらに遅延ゆらぎ，パケットロスの補間も考慮して，3 章で提案したスパイク性遅延変動による長期補間を含めて，統合的に補間を行うハイブリッド補間方式を提案する．ハイブリッド補間方式では，**表 1** に示すように，遅延種別や遅延発生時が有音区間か無音区間か，またはトークスパート区間であるか非トークスパート区間であるか，あるいはパケットロスがあるかにより，従来より提案されている補間方式を併用し，遅延の長さによりスムーズな補間方式の遷移を行う．

4.1 遅延ゆらぎ吸収

定常的に遅延が増減する遅延ゆらぎに対しては，プレイアウトバッファにパケットを蓄積することにより吸収する．この蓄積量は多いとより大きな遅延変動に対応できるが，その反面遅延量が増大して相互通話性能が低下してしまう．そのため，遅延ゆらぎを吸収できる最小値に，蓄積量が制御されることが望ましい．この場合，波形補間を行うことを前提として，音質劣化が認識できない程度のパケット廃棄率になるように蓄積量を決定する手法[1] を用いて制御する．ただし，パケット廃棄が連続する場合は，波形補間しても音質劣化しない範囲となるように蓄積量を制御する．

表 1 音声種別・遅延種別・パケットロス長さによる補間方法
Table 1 Operation differences depending on length of talkspurt, a kind of delay or length of packet loss.

	トークスパート区間		非トークスパート区間	備考
	有音区間	無音区間		
遅延ゆらぎ	プレイアウトバッファで吸収			4.1 節
1〜2 パケットロス	波形補間	周囲雑音補間		4.2 節 4.4 節
1〜2 パケット遅れ	波形補間@			
2 パケットを超えるパケットロス	言い直し補間*		周囲雑音補間*	4.3 節 4.4 節
2 パケットを超えるパケット遅れ（スパイク性遅延）	言い直し補間		周囲雑音補間	
2 パケットを超えるパケット遅れ（スパイク性遅延）＋パケットロス	言い直し補間*		周囲雑音補間*	

@遅れて到着したパケットの廃棄を行う．
*ロスしたパケットの再送を行う．

4.2 有音区間での 1〜2 パケット遅れ・ロス

1〜2 パケットの遅れやロスは，発生頻度も高く，この場合に毎回言い直し補間を行ったのでは，言い直し頻度が多くなり聞きにくくなってしまう．また，1〜2 パケット補間の場合は，波形補間でも十分音質を維持できるので，そのような場合には，従来より提案されている短期の波形補間を行ったほうがよい．この場合は遅れて到着したパケットは廃棄する．短期の波形補間は，使用するコーデックにより推奨されている方法を使用する．本研究で使用した G.711（μ LAW）コーデック[12] では，波形のピッチを検出して，そのサイクルを繰り返すことにより補間を行う方法が推奨されている[6]．また G.723.1（MP-MLQ）コーデック[7] と G.729（CS-ACELP）コーデック[8] では，“フレーム消失の隠蔽” という項目に，仕様として補間方法が含まれている．いずれのコーデックを使用しても，短期の波形補間による音質劣化が十分小さいならば，本研究で使用した G.711（μ LAW）コーデックの場合と変わらない結果が得られることが期待できる．

図 5 (1) に示す音声再生において，パケット 4 が遅れて到着した場合は，図 5 (2) に示すように，パケット 4 を廃棄して波形補間を行う．

4.3 有音区間での 2 パケットを超える遅れ・ロス

スパイク性遅延変動においては，最初はパケットが到着しない時間がどのくらい続くのか分からないため，まず有音区間では波形補間を行い，無音区間では周囲雑音補間を行う．その後，有音区間では，波形補間でも音質が維持できる最大波形補間可能時間を経過しても，次のパケットが到着しない場合は，急激な波形切断による異音が出ないように，波形補間しながら徐々に音量を下げる波形フェードアウト時間に移行する．いったん，波形フェードアウト時間に移行した場合は，次に再生すべきパケットが到着しても，言い直し補間を行う．この後に，完全な無音では違和感があるため，周囲雑音を再生するのであるが，波形フェードアウト時間内にも，周囲雑音を徐々に増大させながら加算させて，波形フェードアウト時間から周囲雑音時間へのつなぎをスムーズにする．周囲雑音時間は，人間の会話で言い直しをする場合でも，時間をおいてから言い直すこともあるため，自然さの点で問題ないことが期待できる．逆に，言い直し補間の前に，ある程度の周囲雑音時間がないと不自然に聞こえるため，最小周囲雑音時間を定め，周囲雑音時間はそれ以上となるようにする．その後，遅れて到着したパケットと時間が連続するように，トークスパートの最初から再生を開始する．再生を開始するタイミングは，次に再生すべき

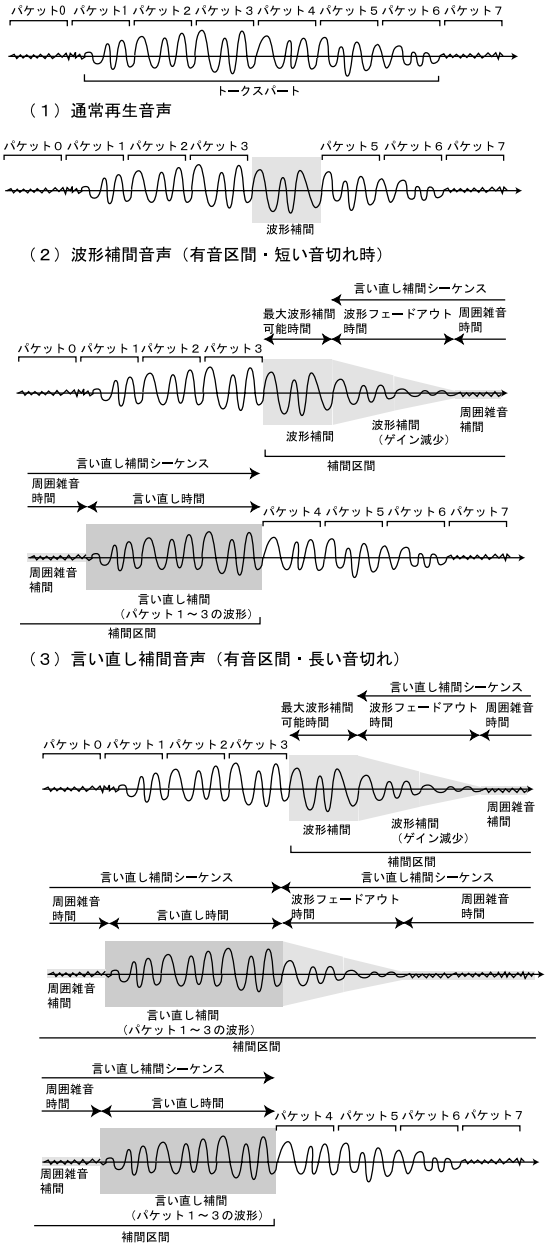


図 5 ハイブリッド補間波形（有声区間）
Fig. 5 Hybrid compensated wave form during talkspurt period.

遅れていたパケットが到着した後であれば安全であるが，それでは遅延時間が増大してしまうため，それよりも前にトークスパートの再生を開始する．(図 5 (3)) この開始タイミングは 5 章で説明する．

会話を言い直し終わっても次のパケットが到着しない場合は，図 5 (4) に示すように，再度波形補間で音声を減衰し，周囲雑音を再生してパケット到着を待ち，

再度言い直し再生をする．

実験的に最大波形補間可能時間を 40 msec，波形フェードアウト時間を 20 msec，最小周囲雑音時間を 180 msec とした．

4.4 無音区間での遅れ・ロス

遅れないで最後に到着した音声パケットが無音の場合は，周囲雑音で補間するのが適切と考えられる．**図 6** (1) に示す音声再生において，パケット 4 が遅れて到着した場合は，図 6 (2) に示すように周囲雑音で補間する．ただしこの場合は，遅れて到着したパケットの廃棄は行わないものとする．理由は，そのパケットの中に，音声が含まれている可能性があるためである．パケットロスしている場合は，図 6 (3) に示すように周囲雑音で補間する．補間可能時間以上の欠損であれば，非トークスパート区間でも再送要求しなくてはならない．理由としては，やはり，ロスしたパケットの中に，音声が含まれていた可能性があるからである．

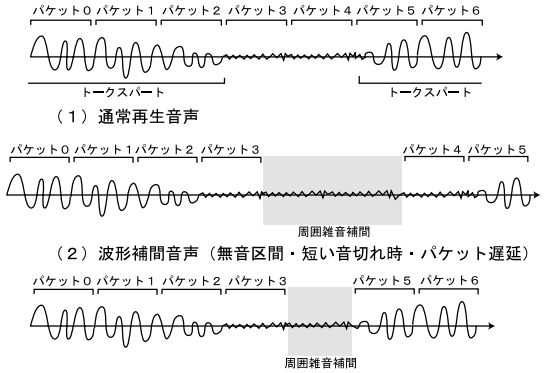


図 6 ハイブリッド補間波形（無音区間）
Fig. 6 Hybrid compensated wave form in silent period.

補間可能時間以上の欠損が起ったとき，トークスパート区間であれば言い直し補間を行う．無音区間でも，有音区間の場合と同じシーケンス動作を行うが，波形フェードアウト時間には，周囲雑音をそのままレベルを変えないで出力する．

5. 言い直し補間のタイミング

図 7 に言い直し補間のタイミングテーブルを示す．音声パケット 0〜2 が，正常に受信側に到着する．その直後にスパイク性遅延変動が起こり，音声パケット 3〜4 はロスし，音声パケット 5〜11 は遅れる．受信側では，次に再生すべきパケットが到着しないため，まず波形補間を行い，次に波形補間を行いながらフェードアウトし，周囲雑音補間を行う．

その後，遅れて到着したパケットと時間が連続するように，トークスパートの最初から再生を開始する．再生を開始するタイミングは，次に再生すべき遅れていたパケットが到着した後であれば安全であるが，それでは遅延時間が増大してしまうため，4.3 節で説明した最大波形補間可能時間，波形フェードアウト時間と最小周囲雑音時間経過後に，次の条件式 (1) を満たしたときにトークスパートの再生を開始する．

$$S \geq P - T \tag{1}$$

S，P，T は図 7 に示されている．S は補間開始からの経過時間で，T は言い直し補間時間長さで，P は予想最大音切れ時間である．この予想最大音切れ時間は，通話開始時の初期値を 1,200 msec とし，これを超えるスパイク性遅延変動による音切れが発生した場合は，予想最大音切れ時間をその時間に書き換える．そ

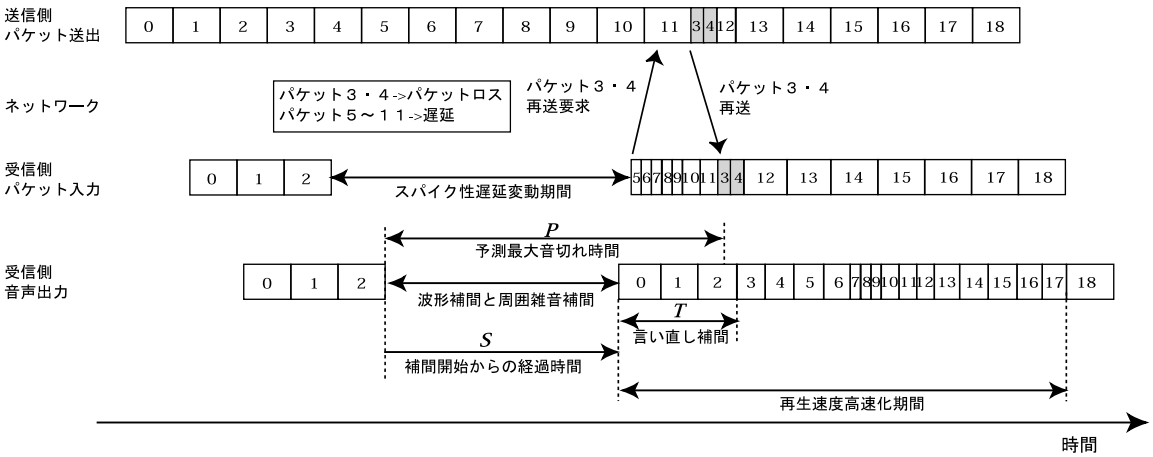


図 7 言い直し補間時の時間テーブル
Fig. 7 Timing of speech repetition compensation.

の後，音切れ時間が予想最大音切れ時間より小さいときは，1,200 msec まで徐々に予想最大音切れ時間を減少させる．初期値の 1,200 msec は，測定した範囲のスパイク性遅延変動の最大値である 800 msec に，後述するパケット再送時間 200 msec と余裕分 200 msec を加えた値である．

次にパケット 5 が到着し，そのとき受信側では，音声パケット 3〜4 がロスしていたことが分かる．受信側では送信側に再送要求を出し，送信側ではそれに応じて音声パケット 3〜4 を再送する．予想最大音切れ時間には，このパケット再送時間も含まれている．言い直し補間によって遅延時間が増大してしまうので，話速変換技術を使用して，速やかに低遅延状態に戻す．この話速変換技術は，有音区間では波形の 1 サイクルを間引くことにより，また無音区間では波形を切り取ることにより，音声のピッチを変化させないで再生速度を速くする技術である[11]．

6. 評価と考察

6.1 評価実験装置

本方式を実装した実験用 IP 電話システムを開発して，言い直し補間の有効性について評価を行った．

PC 上に，ハイブリッド補間を装備した IP 電話機を実装した．この構成を**図 8** に示す．A/D・D/A コンバータを外付けとして，RS232C により PC と接続して遅延時間を抑え，補間評価上，他の音質劣化要因を極力抑えた．パケット送出間隔は 20 msec とし，音声コーデックは G.711（μ LAW）64 kbps とした．スパイク性遅延変動がないときのローカルネットワークでのエンドツーエンド遅延は，80 msec 前後であった．また，評価のため，音声ファイルからの音声を相手に流せるようにした．評価のためのスパイク性遅延変動の発生は，送信側の PC 上でエミュレーションしている．

6.2 言い直し補間の評価方法

VoIP の音質評価は，遅延やエコーなども含めた総合的通話品質評価である R 値[13] や，その中の音質評価で使用されている客観的音質評価尺度である，PESQ (Peceptual evaluation of speech quality)[14] を用いるのが一般的である．

しかしながら，PESQ など客観的音質評価法は，原音波形と受信音声波形との比較である．PESQ の場合は，遅延変動やパケット損失により極端に評価が下がることのないような処理が入っているが，言い直し補間では波形的には大幅に原音と異なってしまうため，PESQ では言い直し補間の音質評価はできない．そのため主観的評価方法により評価を行うこととした．この場合も，MOS（Mean Opinion Score）のように，音質を 5 段階で評価する方法では，言い直し補間に対しては評価基準が不明確になるので，相手に聞き直す必要があるかどうかで評価を行う方法を考案し，評価を行った．

また，音質と同様に重要な評価項目である遅延については，通常の最大あるいは平均遅延測定では，言い直し補間を行ったときのように，大きく遅延が変動することを考慮していないため，正確な評価ができない．そのため話者交代時点での遅延確率を測定する方法を考案し，評価を行った．

6.3 言い直し補間の音質評価

1 つの実験用 IP 電話端末より，音声ファイルに記録されているサンプル音声をもう 1 つの実験用 IP 電話端末に送り，そこで受信された音声を音声ファイルに記録した．そのとき，8 秒間隔の一定周期で，800 msec の大きさのスパイク性遅延変動を与えておいたが，音声ファイルに記録された音声が，たとえば文の最初の方にだけ偏って音声欠損が起こっているというようなことがないことを確認しておいた．また話速度高速化率は，評価に用いた音声で，話速度高速化により著しく認識率が下がることのない最大高速化率を実験的に選んだ結果，トークスパート区間 20%，非トークスパート区間 50%とした．46 人の被験者に，次の 3 種類の音声途切れ処理方法を使用した場合の，受信側で記録されたサンプル音声を，それぞれ 2 分間ずつ聞いてもらい評価を行った．

(1) 遅れて到着したパケットを廃棄する場合（図 2 (2)）
(2) 遅れて到着したパケットを廃棄しないで再生する場合（図 2 (3)）
(3) 言い直し補間を行う場合（図 2 (4)）

被験者にそれぞれのサンプル音声で，次の（A）と（B）の場合がそれぞれ何回あったか記録してもらった．ただし同じ箇所で（A）と（B）は同時に記録しないようにしてもらった．

（A） 意味は分かったが，たぶん相手に同じこと

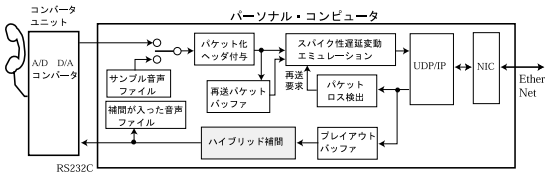


図 8 ハイブリッド補間実験用 IP 電話端末構成
Fig. 8 Block diagram of IP telephony system.

表 2 言い直し補間の評価結果
Table 2 Result of quality evaluation of speech repetition compensation.

	$P(A)$	$P(B)$	$P(A)+P(B)$	備考
遅れて到着したパケットを廃棄する場合	24%	52%	76%	図 2 (2)
遅れて到着したパケットを廃棄しないで再生する場合	29%	18%	47%	図 2 (3)
言い直し補間	11%	3%	14%	図 2 (4)

$P(A)$ 意味は分かったが，たぶん相手に同じことをもう一度言うように頼むだろう確率.
$P(B)$ 意味がまったく分からなかった，当然相手に同じことをもう一度言うように頼む確率.
被験者 10 代〜60 代 日本人 男女 46 人 日本語
スパイク性遅延変動 8 秒間隔一定周期 800 msec の大きさ
サンプル音声 2 分間 “多言語音声データベース 1994”[15]
（無音部分を短縮して使用した）
話速度高速化率 トークスパート区間 20%
非トークスパート区間 50%

をもう一度言うように頼むだろう.

(B) 意味がまったく分からなかった．当然相手に同じことをもう一度言うように頼む.

（A）の発生確率 $P(A)$ を式 (2) により，(B) の発生確率 $P(B)$ を式 (3) により求めた.

$$P(A) = \overline{A}/Sp \tag{2}$$

$$P(B) = \overline{B}/Sp \tag{3}$$

$\overline{A}$ は 46 人の被験者により記録された（A）の平均回数であり，同様に $\overline{B}$ は記録された（B）の平均回数である．Sp は 2 分間のスパイク性遅延変動の発生回数であり 15 である.

$P(A)$，$P(B)$ と $P(A)+P(B)$ の結果を**表 2** に示す．ここで $P(A)+P(B)$ は相手に同じことをもう一度言うように頼む確率を示している．言い直し補間を行った場合には，$P(A)+P(B)$ が 14%と，他の方式に比較して大幅に減少しており，言い直し補間の有用性が確認された.

6.4 言い直し補間時の遅延増大評価

言い直し補間を行うと，スパイク性遅延変動による遅延増大よりも大きな遅延が一時的に発生する．**図 9** に，会話を行っているときに，10 秒ごと 800 msec のスパイク性遅延変動が起こったときの音声の遅延変動を調査するために，プレイアウトバッファに保持されているパケットの滞留時間の変化を計測したものを示す．これによると，スパイク性遅延変動が起こったときには，スパイクの大きさ以上の遅延が発生している．その後遅延は，話速度高速化により急速に減少している.

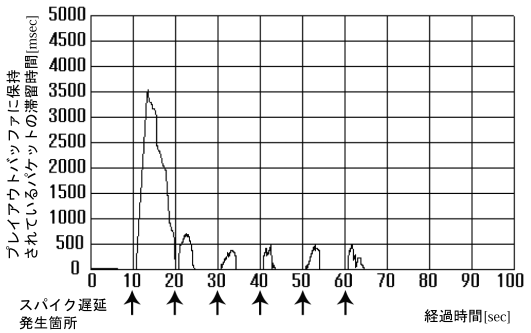


図 9 スパイク性遅延変動が起こったときのプレイアウトバッファに保持されているパケットの滞留時間の例
Fig. 9 Example of delay fluctuation due to the number of packets stored in playout buffer when a delay spike occurs.

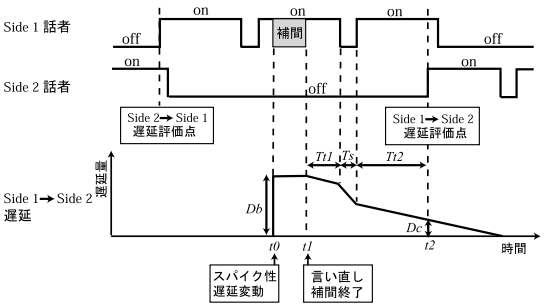


図 10 言い直し補間の会話への影響の評価方法
Fig. 10 Evaluation method for influence of speech repetition compensation over dialog.

6.5 会話への言い直し補間の影響

遅延が会話にどうような影響があるか評価する場合は，**図 10** で示すように，話者が交代する時点での遅延で評価するべきである．理由は，一方の話者が話し続けており，その間もう一方の話者が返事さえしない場合は，遅延は何も会話に影響を与えないからである.

最初に，被験者 2 人に実験用 IP 電話端末で通常会話を行ってもらい，いつトークスパートが始まり，いつ終わったかというトークスパートログを，両話者について記録した．そして，スパイク性遅延変動は音声と関係なく一定の確率で発生するので，スパイク性遅延変動がある時点で起こったとした場合に，言い直し補間を行ったときの，話者が交代する時点での遅延を式 (4) により求めた.

$$Dc = Db - Tt \times St - R \tag{4}$$

ここで，Dc は話者が交代する時点 $t2$ での遅延である．Db は言い直し補間終了時刻 $t1$ での遅延である．ここでは，スパイク性遅延変動発生時刻 $t0$ で，図 5 (3)

に示した最大波形補間可能時間，波形フェードアウト時間，周囲雑音時間と話速度高速化を行った後の言い直し補間の時間分，再生時間が一度に戻り，遅延が時刻 $t0$ で一度に増大したと考えている．なお，この最小値は予想最大音切れ時間となる．ただし，非トークスパート区間でスパイク性遅延変動が起こった場合は，周囲雑音補間を行うので，Db は周囲雑音補間終了時刻 $t1$ での遅延となり，遅延量はスパイク性遅延変動の大きさと等しくなる．時刻 $t0$ から時刻 $t1$ までは遅延は一定であり，時刻 $t1$ 以降に，話速度高速化により遅延が減少すると考える．Tt は時刻 $t1$ と $t2$ の間で，トークスパート区間の時間長の総和である．たとえば，図 10 の場合には，Tt は式 (5) のようになる．

$$Tt = Tt1 + Tt2 \tag{5}$$

St はトークスパート区間での話速度高速化率である．R は時刻 $t1$ と $t2$ の間の，非トークスパート区間での遅延短縮時間である．非トークスパート区間においても，トークスパート区間と同じように話速度を高速化する方法を使用した場合は，R は式 (6) によって求められる．

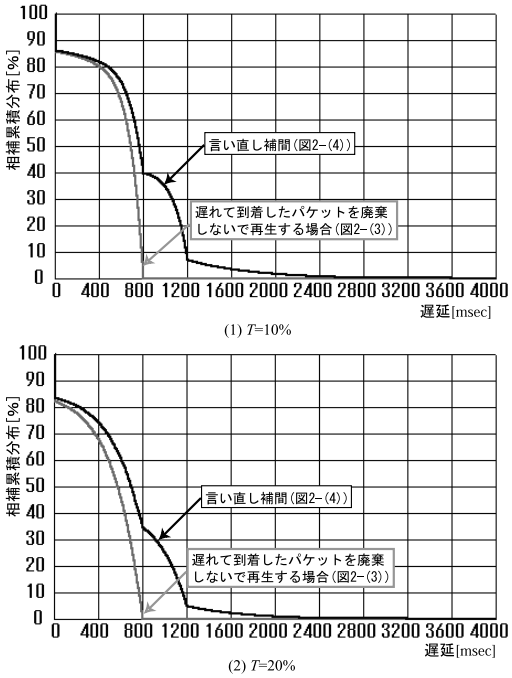
$$R = Ts \times Ss \tag{6}$$

ここで Ss は非トークスパート区間での話速度高速化率であり，Ts は時刻 $t1$ と $t2$ の間での，非トークスパート区間の時間長の総和である．この方法は，6.3 節の音質評価で使用した．しかし，この方法は，文と文の間の時間がほぼ一定であるサンプル音声では問題なかったが，一般通話では，文と文の間の時間が短くなりすぎて不自然になることがあった．これを解決するために，非トークスパート区間では，時間短縮する際に最小時間を設定することとした．この自然さを維持するための最小時間は，文の途中では 200 msec，文と文の間では 1,000 msec であると報告されている[11]．しかし文の途中か文と文の間かを厳密に判定することは難しいので，文と文の間の最小時間の 1,000 msec を採用し，3.3 節で説明したように，トークスパート区間の最後 100 msec は無音区間であることを考慮し，非トークスパート区間が 900 msec を超えたときのみ，それ以降を 100%の話速度高速化率で高速化し，900 msec 以下のときは高速化しないこととした．これはすなわち，900 msec を超えるときに 900 msec に短縮するということである．この方法を使用した場合の R は，時刻 $t1$ と $t2$ の間で，900 msec を超える非トークスパート区間の 900 msec 超過分の総和となる．

最後に，1 msec ごとに遅延 Dc を計算し，相補累積分布を計算した．

40 人の被験者で，通話を行う被験者の組合せを変えた 2 分間の通話を 25 回行い，50 のトークスパートログを記録した．遅延 Dc の相補累積分布をそれぞれのログより計算し，その平均を求めた．この結果を**図 11** に示す．この遅延時間の相補累積分布は，ある遅延よりも大きい遅延の確率を示している．たとえば，図 11 (1) において，言い直し補間の場合の 1,200 msec より大きい遅延の確率は約 7%である．話速度高速化により認識率が著しく低下する場合には，話速度高速化率を小さくする必要がある．そのため，図 11 にトークスパート区間話速度高速化率が異なる 2 種類の結果を示した．これらを比較すると，予想最大音切れ時間の大きさである 1,200 msec 以上の遅延確率は，図 11 (1) が約 7%で図 11 (2) が約 4%であり，問題となる可能性のある大きな遅延の確率はあまり変わらないと考えられる．そのため，実使用では話速度高速化による認識率低下の危険を避けるため，トークスパート区間話速度高速化率は，10%程度に小さくした方がよいと考えられる．

言い直し補間は他の方式に比較し遅延を増大させる．



T トークスパート区間話速度高速化率
予想最大音切れ時間 1200msec
遅延増大時 900msec以上の非トークスパート区間を900msecに短縮
スパイク性遅延変動の大きさ 800msec
被験者 10代～60代 日本人 男女40人 日本語
試験方法 被験者の組み合わせを変えて2分間通話を25回行った時の平均

図 11 話者の交代時点での遅延時間の相補累積分布
Fig. 11 Delay complementary cumulative distribution at the point of change of speaker.

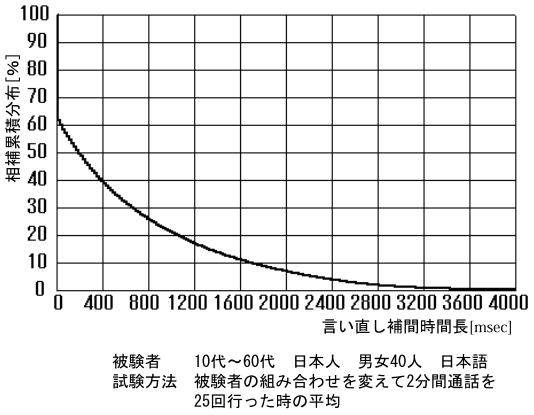


図 12 言い直し補間時間長の相補累積分布
Fig. 12 Complementary cumulative distribution of speech repetition time duration.

筆者らは，言い直し補間を使用して通話を行った．このときに，16 秒に 1 回 800 msec の大きさのスパイク性遅延変動を与え，話速度高速化率はトークスパート区間 10%とし，非トークスパート区間では 900 msec を超えるときに 900 msec に短縮するようにしておいたが，遅延を意識することはなかった．この理由としては以下のことが考えられる．予想最大音切れ時間より大きな遅延は，図 11 (1) によると約 7%であり，このように大きな遅延の頻度が少ないならば，たまの遅延増大は，相互通話性能を維持するためには許されると考えられる．人間も時々遅く返事をすることがあり，その場合，相手の返事を待つからである．

6.6 言い直し補間時間長の評価

6.5 節で使用した 50 のトークスパートログを使用して，**図 12** に示すような言い直し補間時間長の相補累積分布を計算した．

言い直し補間時間長が長すぎると，聞いている人は不快感を感じることが考えられるので，3 秒以上の言い直し補間は避けるべきと考えられる．これはトークスパート保持バッファに補間されている音声パケットが，3 秒分以上になったとき，言い直し補間を行わないで，遅れて到着したパケットを廃棄しないで再生する方式（図 2 (3)）に切り替えることにより可能となる．この方法を使用すると，図 12 より 3 秒以上の言い直し補間時間長確率は 1%程度なので，スパイク性遅延変動が 60 秒に 1 回起こるとしたときには，音声欠損は 1 時間に 1 回程度になる．

7. 結　　論

IP ネットワークのスパイク性遅延変動に対する VoIP の音切れ補間の方式として，言い直し補間を提案し実装，評価を行った．その結果，言い直し補間の有効性について確認した．また，実用化のため，遅延とパケットロスの状況により補間方法を変えるハイブリッド補間方式を提案した．

言い直し補間を行うと，補間後に遅延が増大する．そのため，話速変換技術を導入し，これにより話速度を上げて遅延をすぐに元の低遅延状態にもどした．これにより遅延増大を一時的なものにした．そのため，実際の会話では遅延を意識することはなかった．

謝辞 本研究の機会を与えてくださった檜山竹生（株）エイビット代表取締役社長に感謝いたします．

参 考 文 献

1) 星　徹，谷川桂子，松井　進，岩見直子，寺田松昭：LAN 環境における負荷適応制御を用いた低遅延リアルタイム音声通信システム，情報処理学会論文誌，Vol.40, No.7, pp.3063–3073 (1999).
2) Markopoulou, A.P., Tobagi, F.A. and Karam, M.J.: Assessing the Quality of Voice Communications Over Internet Backbones, IEEE/ACM Trans. Networking, Vol.11, No.5 (Oct. 2003).
3) Liu, F., Kim, J.W. and Kuo, C.C.J.: Adaptive delay concealment for internet voice applications with packet-based time-scale modification, *Proc. International Conference on Acoustics Speech and Signal Processing ICASSP*, Salt Lake City (May 2001).
4) Liang, Y.J., Farber, N. and Girod, B.: Adaptive Playout Scheduling and Loss Concealment for Voice Communication over IP Networks, *IEEE Trans. Multimedia* (Apr. 2001).
5) 青木直史，大宮尚弘，中野隆司，小牧憲子，山本　強，青木由直：VoIP におけるステガノグラフィを用いたパケット損失の一隠蔽法，信学会総合大会，東北大学 (Mar. 2003).
6) ITU-T Recommendation G.711-Appendix I: A high quality low-complexity algorithm for packet loss concealment with G.711 (Sep.1999).
7) ITU Recommendation G.723.1: Dual-rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s (Mar. 1996).
8) ITU-T Recommendation G.729: Coding of speech at 8 kbit/s using conjugate structure-algebraic code excited linear prediction (CS-ACELP) (Mar. 1996).
9) Le, L., Sanneck, H., Carle, G. and Hoshi, T.: Active Concealment for Internet Speech Transmission, Lecture Notes in Computer Science 1942, Active Networks, *Proc. 2nd International Working Conference*, IWAN 2000 Tokyo, Japan, Oct. 2000, Yasuda, H. (Ed.), pp.239–248, Springer-Verlag, Berlin, Heidel-

berg (2000).
10) 船越孝太郎，徳永健伸：話し言葉における言い直しの処理，情報処理，Vol.45, No.10, pp.1032–1037 (2004).
11) 今井 篤，池沢 龍，清山信正，中村 章，都木徹，宮坂栄一，中林克己：ニュース音声を対象にした時間遅れを蓄積しない適応型話速変換方式，信学論（A），Vol.J83-A, No.8, pp.935–945 (2000).
12) ITU-T Recommendation G.711: Pulse Code Modulation (PCM) of Voice Frequencies (1972).
13) TTC 標準 JJ-201.01, IP 電話の通話品質評価法 (Apr. 2003).
14) ITU-T Recommendation P.862: Peceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow band telephone networks and speech codecs (Feb. 2001).
15) NTT アドバンステクノロジ（株）：多言語音声データベース 1994 (Feb. 2006). http://www.ntt-at.co.jp/product/denwa11/index.html

(平成 17 年 11 月 16 日受付)
(平成 18 年 5 月 9 日採録)

小日向 肇（学生会員）

昭和 58 年東京農工大学電子工学科卒業．同年（株）ナカミチ入社．平成 2 年（株）ソニー入社．平成 3 年（株）エイビット入社（現職），平成 6 年より VoIP の研究開発に従事．現在，東京工科大学博士後期課程在学中．電子情報通信学会会員．

星 徹（フェロー）

昭和 44 年東京工業大学工学部電気工学科卒業．同年（株）日立製作所入社．平成 15 年 3 月まで戸塚工場，システム開発研究所，中央研究所等で，交換システム，PBX，マルチメディア LAN，CSCW，デスクトップ会議，CTI，IP テレフォニー等の研究開発に従事．この間，昭和 50 年カリフォルニア大学ロスアンジェルス校（UCLA）大学院コンピュータサイエンス専攻修了．平成 15 年 4 月より東京工科大学コンピュータサイエンス学部教授（現職）．マルチメディアコミュニケーション，RFID タグ応用等，ユビキタスネットワーキングの研究に従事．工学博士．平成 13〜17 年まで情報処理学会グループウェアとネットワークサービス研究会主査．電子情報通信学会，IEEE，ACM 各会員．平成 18 年 3 月情報処理学会フェロー．

松下 温（フェロー）

昭和 38 年慶應義塾大学工学部電気工学科卒業．昭和 43 年イリノイ大学大学院コンピュータサイエンス専攻修了．平成 1 年より平成 14 年 3 月まで慶應義塾大学理工学部教授．平成 14 年 4 月より東京工科大学教授および慶應義塾大学理工学部客員教授．工学博士．マルチメディア通信，コンピュータネットワーク，グループウエア等の研究に従事．情報処理学会理事，同学会副会長，マルチメディア通信と分散処理研究会委員長，グループウエア研究会委員長，電子情報通信学会，情報ネットワーク研究会委員長，MIS 研究会委員長，バーチャルリアリティ学会，サイバースペースと仮想都市研究会委員長等を歴任．現在，情報処理学会 ITS 研究会委員長，郵政省，通産省，建設省，農水省，都市基盤整備公団，行政情報システム研究所等の委員長，座長，委員を多数歴任．『やさしい LAN の知識』（オーム社），『201x 年の世界』（共立出版）等著書多数．平成 5 年情報処理学会ベストオーサ賞，平成 7 年および平成 12 年情報処理学会論文賞，平成 12 年 10 月 20 日情報処理学会 40 周年記念 90 年代学会誌論文賞，平成 12 年 10 月 2 日電子情報通信学会フェロー，平成 12 年 10 月 VR 学会サイバースペース研究賞，平成 13 年 5 月情報処理学会功績賞，平成 14 年 3 月情報処理学会フェロー．電子情報通信学会，人工知能学会，ファジイ学会，IEEE，ACM 各会員．