統計的手法を用いた音声信号の復元手法の改良

黒岩 眞吾[1], 柘植 覚[1], 任 福継[1], 來山 征士[1]

A Novel Packet Loss Concealment Algorithm based on Satistical Methods

by

Shingo Kuroiwa, Satoru Tsuge, Fuji Ren, Seishi Kitayama

In recent years, IP telephone use has spread rapidly thanks to the development of VoIP (Voice over IP) technology. However, an unavoidable problem of the IP telephone is deterioration of speech due to packet loss, which often occurs on the wireless network. To overcome this problem, we propose a novel packet loss concealment algorithm using speech recognition and synthesis. This proposed method uses linguistic information and can deal with the lack of syllable units which conventional methods are unable to handle. We conducted subjective and objective evaluation experiments. These results showed the effectiveness of the proposed method. Although there is a processing delay in the proposed method, we believe that this method will open up new applications for speech recognition and speech synthesis technology.

Key words: Packet Loss Concealment, Missing Feature Theory, Speech Recognition, Speech Synthesis

1. はじめに

近年，DSL (Digital Subscriber Line)，ケーブルインターネット等によるアクセス回線のブロードバンド化，IP (Internet Protocol) ネットワーク関連機器の高機能化などによって，IP ネットワーク上に展開される様々なアプリケーションを多くの人が利用できる環境が整備されつつある。特に，IP ネットワーク上で音声を取り扱うことを可能とする VoIP (Voice over IP) 技術の発展によって，音声電話を IP ネットワーク上で提供するサービスが現れ，これまでの電話網における回線交換を中心とするネットワークからパケット交換による IP ネットワークに向かう変化が急速に進んでいる。このように，IP ネットワーク技術を利用して提供される電話サービスのことを IP 電話と呼ぶ。

IP 電話の本質的な問題として (特に無線区間がある場合)，音声パケットがネットワークの途中で消失するパケット損失や転送速度及び到着間隔がばらつく揺らぎが挙げられる。これらの問題により，音声を再生するべきタイミングでパケットを受け取れない場合には，その部分の音声が途切れたり，雑音が発生し，著しく音声品質を劣化させる。パケット損失や揺らぎによる影響を軽減させ，高い音声品質を維持するためには，パケット損失補償 (Packet Loss Concealment) が重要となる。パケット損失補償とはパケット損失区間を違和感の少ない信号で埋める技術であり，その方法は音声の符号化方式によって異なる。例えば，IP 電話で一般的に利用されている G.711 の場合には，バッファに過去の音声パケットを保持しており，保持している過去のデータを分析し，減衰させながら繰り返すことでパケット損失区間を補償している [1]。しかし，この手法は最大で 5 フレーム (50ms) までのパケット損失しか補償できない。さらに，2 音素以上の欠落には原理的に対応していないという問題がある。

本論文では，より長い区間の音声途切れ及び複数音素の欠落に対応するために，統計的手法を用いた音声認識と音声合成を利用した音声途切れ補間手法を提案する。提案手法は，(1) Missing Feature Theory に基づく音声認識により，途切れ区間の音素片列を前後の言語情報及び音響情報から推定し，(2) 推定した音素片列に基づき HMM 音声合

1 徳島大学 工学部 知能情報工学科
Deparatment of Information Science & Inteligent Systems, Faculty of Engineering, The University of Tokushima
連絡先: 〒 770-8506 徳島市南常三島町 2-1

成を行い，途切れ区間及び前後の音声波形を生成し補間する。これにより，より長い区間の音声途切れ及び複数音素の欠落に対応することが可能となる。本論文ではリアルタイム処理を行うまでには致っていないが，将来的に提案手法がパケット損失補償の有効な一手法となり得ることを示す。

2 では提案手法である統計的手法を用いた音声途切れ補間手法について述べる。**3** では音声途切れ区間の音素片列を推定する際に用いる Missing Feature Theory について述べる。**4** では補間に用いる音声信号を発声者に適応するための話者適応手法について述べる。**5** では提案手法の有効性を検証するために，主観評価実験及び客観評価実験を行い，その結果について考察する。最後に **6** において本稿のまとめを述べる。

2. 音声認識・音声合成を用いた音声途切れ補間手法

2.1 概　　要

図 1 に提案手法のブロック図を示す。提案手法は，(1) 途切れ区間のある音声に対して，途切れ区間の音素片 (HMM 状態) を推定するために Missing Feature Theory [2], [3] に基づく音声認識を行い，途切れ区間及びその前後に対応する音素片列を推定する。(2) 推定した音素片列に対し， 尤度最大化基準に基づくパラメータ生成手法 [4] を用い，メルケプストラム列を生成する。このとき，途切れ区間以外の音声から話者の特徴を抽出しパラメータ生成に用いる音響モデル (HMM) を予め話者適応しておく。また，途切れ区間の有声音のピッチ情報については，途切れ区間前後のピッチ情報を用いて，線形補間により求める。その後，メルケプストラム列と補間したピッチ情報を用いて，MLSA (Mel Log Spectral Approximation) フィルタ [5] で音声合成を行う。最後に，生成した合成音声のうち，途切れ区間に対応する部分を用いて，途切れ区間を補間する。

2.2 定　式　化

以下に，音声途切れ補間手法の基本的な概念を説明する。あるフレーム化された音声信号を

$$\boldsymbol{X} = \{x(1), x(2), \ldots, x(i), \ldots, x(i+n), \ldots, x(N)\}$$

とする。ここで，N は音声信号の総フレーム数である。この音声信号のうち

$$\boldsymbol{X_m} = \{x(i), \ldots, x(i+n)\}$$

がパケット損失により失われたとする。この時，以下の条件付き確率

$$P(\{\tilde{x}(i), \ldots, \tilde{x}(i+n)\} | \{x(1), \ldots, x(i-1)\}, \{x(i+n+1), \ldots, x(N)\}) \quad (1)$$

を最大とするような，

$$\tilde{\boldsymbol{X}}_m = \{\tilde{x}(i), \ldots, \tilde{x}(i+n)\} \quad (2)$$

を求めて下記のように音声信号を補間する。

$$\{x(1), \ldots, x(i-1), \tilde{x}(i), \ldots, \tilde{x}(i+n), x(i+n+1), \ldots, x(N)\} \quad (3)$$

これらは，失われていない音声信号からパケット損失区間の音声信号を推定し，作成することで，パケット損失を補償することを表している。

式 (1) の替わりに，

$$P(\{\tilde{x}(i), \ldots, \tilde{x}(i+n)\} | \{x(1), \ldots, x(i-1)\}) \quad (4)$$

を用いて左コンテキストのみから音声信号を予測すれば，リアルタイム処理が可能となる。また，若干のディレイ ($i+M$ フレーム) を許容し，下記の確率を最大にする方法も考えられる。

$$P(\{\tilde{x}(i), \ldots, \tilde{x}(i+n)\} | \{x(1), \ldots, x(i-1)\}, \{x(i+n+1), \ldots, x(i+n+M)\}) \quad (5)$$

これら方法のうち，本論文では手法の基礎的検討を目的とすることから，以下の章では式 (1) を用いる。

2.3 音声認識・音声合成を用いた補間手法

直接，式 (1) を最大化することは困難である。そこで，音声認識・音声合成を用いて式 (1) を最大化する $\tilde{\boldsymbol{X}}_m$ を求める。

（1） 下記の式を最大化する $\{\tilde{s}(1), \ldots, \tilde{s}(L)\}$ を求める。

$$P(\{\tilde{s}(1), \ldots, \tilde{s}(L)\} | \{x(1), \ldots, x(i-1)\}, \{x(i+n+1), \ldots, x(N)\}) \quad (6)$$

ここで，$\tilde{s}(i)$ は，音声信号 $\{x(1), \ldots, x(i-1)\}, \{x(i+n+1), \ldots, x(N)\}$ が与えられたときの，消失区間 (i〜$i+n$) を含む音声認識結果 (単語系列，音素系列，状態系列，分布系列のいずれか，$L \leqq N$) である。本論文では Missing Feature Theory を適用した音声認識により状態時系列を求めた。Missing Feature Theory の詳細については **4** で述べる。

（2） (i) で求めた $\tilde{s}(i)$ を時刻に対応付ける。通常音声認識の過程で (i)(ii) は同時に行われる。求めた時系列を $\{\hat{s}(1), \ldots, \hat{s}(N)\}$ とする。

（3） 下記の式を最大にする $\{\tilde{x}(i), \ldots, \tilde{x}(i+n)\}$ を求める。

$$P(\{\tilde{x}(i), \ldots, \tilde{x}(i+n)\} | \{\hat{s}(1), \ldots, \hat{s}(N)\}) \quad (7)$$

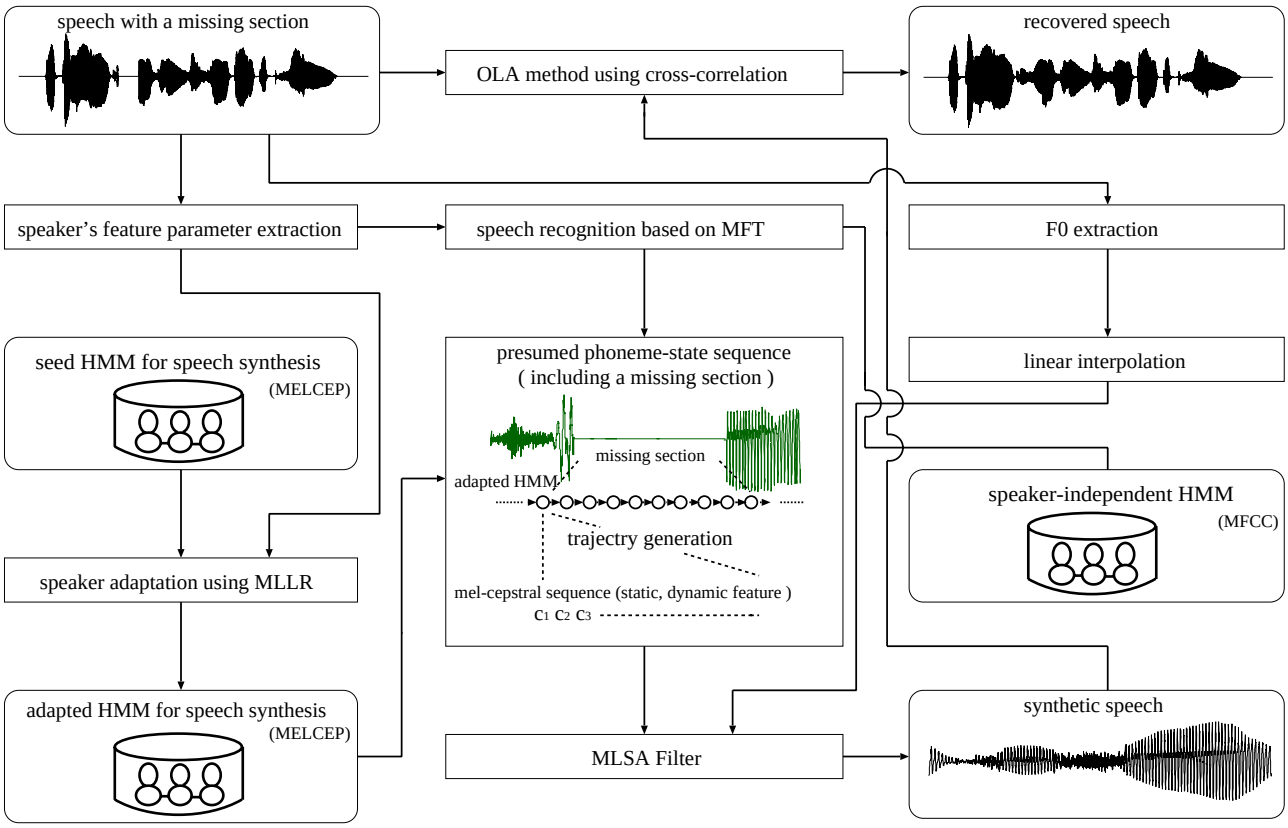

図 1　提案手法のブロック図

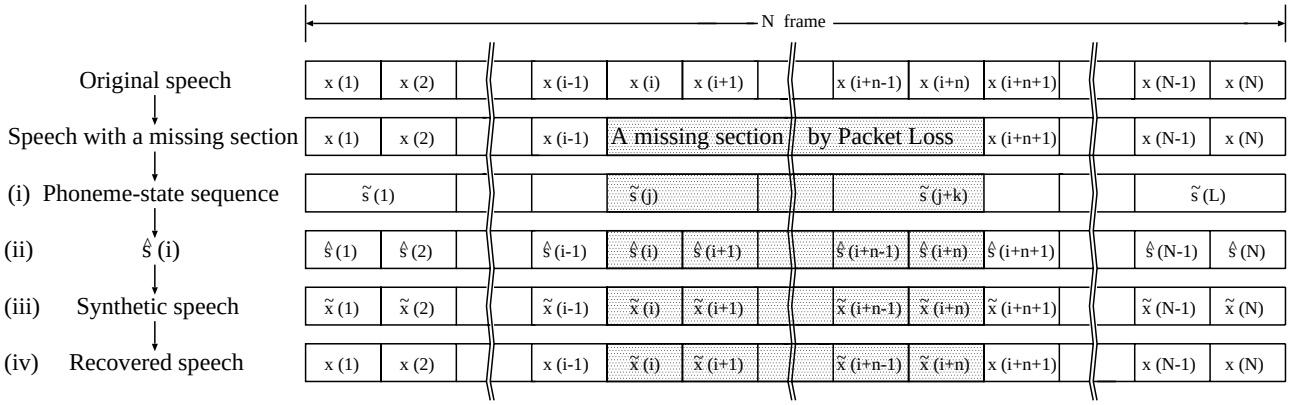

図 2　提案手法によるデータ補間の過程

式 (7) では，$\{\hat{s}(1), \ldots, \hat{s}(N)\}$ の全てを考慮して消失区間の音声信号を生成しているが，実際には消失区間の前後数フレームから音声信号を生成できる。この際，音声信号の生成には HMM 音声合成を用いる。

（4） 以上により下記の復元・補間された音声信号が生成される。

$$\{x(1), \ldots, x(i-1), \tilde{x}(i), \ldots, \tilde{x}(i+n), \\ x(i+n+1), \ldots, x(N)\} \qquad (8)$$

図 2 に上記の処理によりデータが補間される過程を示す。図中の (i)〜(iv) は上記に示した処理項目に各々対応する。

3. Missing Feature Theory を用いた途切れた音声の認識

3.1 概　　念

Missing Feature Theory [2], [3] とは，失われた情報は使わずに，残された情報のみで認識を行う手法である。以下，文献 [2] に沿って，Missing Feature Theory の手法の説明を行う。本手法は文献 [3] における Marginalization にあたる。

表 1　音響分析条件 (音声認識)

sampling rate	16 kHz
frame length	25 ms
frame shift	10 ms
window	Hamming
feature vector	1–12 MFCCs (CMS), Δ MFCCs, Δ LogPower (total 25)

混合正規分布 HMM による尤度計算において，ある時刻の D 次元の出力ベクトル $\boldsymbol{x}$ に対する，状態 c の出力尤度 $p(\boldsymbol{x}|c)$ は，

$$p(\boldsymbol{x}|c) = \sum_{j=1}^{M} w_j \prod_{i=1}^{D} N(u_{ij}, \sigma_{ij}^2) \tag{9}$$

と表すことができる。ただし，M は混合数，w_j は第 j 分布の重み係数，$N(u_{ij}, \sigma_{ij}^2)$ は$\boldsymbol{x}$の第 i 次元要素に対する第 j 分布の出力尤度を各々表す。ここで，出力ベクトル $\boldsymbol{x}$ のいくつかの要素が失われた場合を考えるために，式 (9) を式 (10) に示すように存在する要素 ($present$) と失われた要素 ($missing$) に分解する。

$$p(\boldsymbol{x}|c) = \sum_{j=1}^{M} w_j \prod_{i \atop present} N(u_{ij}, \sigma_{ij}^2) \prod_{i \atop missing} N(u_{ij}, \sigma_{ij}^2) \tag{10}$$

$p(\boldsymbol{x}_p|c)$ は，失われた要素のベクトル $\boldsymbol{x}_m$ で $p(\boldsymbol{x}|c)$ の全空間積分をとったものと考えることができる。これは，式 (10) の $missing$ の項を 1 (unity) とすることに等しく，$p(\boldsymbol{x}_p|c)$ は単純に次の式で表現できる。

$$p(\boldsymbol{x}_p|c) = \sum_{j=1}^{M} w_j \prod_{i \atop present} N(u_{ij}, \sigma_{ij}^2) \tag{11}$$

3.2　音声途切れへの適用

本論文では前節で示した手法を，音声途切れに適用するにあたり，パケットロスが生じたフレームは，ベクトルの要素がすべて失われたと考えた。すなわち，当該フレームの出力尤度は，すべての状態で等尤度とし，遷移確率のみを用いて探索を行った。また，音声途切れに対して，Missing Feature Theory を適用することが有効であることを示すために以下の実験を行った。

3.3　音声認識実験

評価データには FAK が発声した新聞記事読み上げ文 100 発声を用いた。各発声に対して，音声途切れの開始位置を 1.0 (s) とし，そこから様々な区間 (0.05 ～ 0.5 (s)) の音声途切れを発生させ Missing Feature Theory 適用の有無による単語正解精度 (word accuracy) の比較を行った。このとき，音声途切れ区間は既知としている。これは実際の IP 電話では送信するパケットに，再生するパケット順序をシーケンス番号として記録しており，IP ネットワーク内で損失したパケットを受信側で知ることが可能なためである。デコーダには大語彙連続音声認識エンジン julius3.3p4 [6] を，音響モデルは女性話者 133 名 が発声した音素バランス文 20,958 発声を用いて学習した状態共有 triphone HMM (2000 状態，16 混合) [7] を使用した。表 1 に音響分析条件を，表 2 に音素ラベルを示す。

表 2　音素ラベルの種類

vowels	a, i, u, e, o
long vowels	a:, i:, u:, e:, o:
consonants	b, d, gy, my, py, sh, ry, z, by, dy, n, w, ts, ky, g, hy, j, m, ny, ch, r, y, p, t, k, f, h, s
choked sound	q
syllabic nasal	N
silence	silB, silE, sp

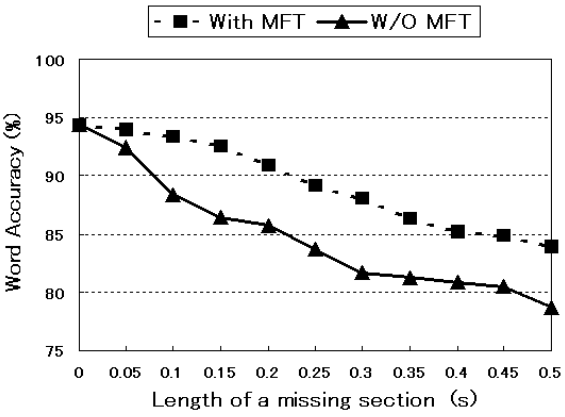

図 3　単語正解精度 (音声途切れ開始位置 1.0 (s) の場合)

また，言語モデルや辞書は音声認識システム [7] の付録 CD-ROM に収録されている 2 万語彙の辞書及びバイグラム，トライグラムを使用した。また，単語正解精度は以下の式で計算した。

$$\text{単語正解精度 (\%)} = \frac{N - D - S - I}{N} \times 100 \tag{12}$$

ここで，N は正解系列に含まれる全単語数，D は脱落誤り数，S は置換誤り数，I は挿入誤りをそれぞれ表している。

3.4　実験結果・考察

音声途切れ開始位置が 1.0 (s) の場合の認識結果を図 3 に

示す。点線は Missing Feature Theory を適用した場合の単語正解精度を示しており，実線は Missing Feature Theory を適用しない場合の単語正解精度を示している。これらの結果から，Missing Feature Theory を適用することで認識率の低下を抑制し，音声途切れ区間の音素片をより正しく推定できることがわかる。

4. MLLR を用いた話者適応

提案手法を不特定多数の話者に適用するため，合成用の音響モデルを最尤回帰 (MLLR: Maximum Likelihood Linear Regression) 法 [8] を用い話者適応し音声合成に用いる。本章では MLLR について文献 [8] に従い説明すると共に，予備実験結果について述べる。

4.1 MLLR

MLLR 法は，音響特徴空間における話者間の線形写像を用いる話者適用手法であり，取扱い易さと性能の高さから音声認識の分野において広く用いられている．

MLLR 法では，次式で表されるように，HMM のガウス分布の平均ベクトル $\hat{\boldsymbol{\mu}}$ は適応前の平均ベクトル $\boldsymbol{\mu}$ のアフィン変換によって与えられる．

$$\hat{\boldsymbol{\mu}} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} \tag{13}$$

ここで $\boldsymbol{A}$ は $n \times n$ の行列，$\boldsymbol{b}$ は次元数 n のベクトル，n は特徴ベクトルの次元数である．式 (12) は，次式のように線形変換に書き直すことができる．

$$\hat{\boldsymbol{\mu}} = \boldsymbol{H}\boldsymbol{\xi} \tag{14}$$

ここで $\boldsymbol{H}$ は適応のための $n \times (n+1)$ の変換行列，$\boldsymbol{\xi}$ は，

$$\boldsymbol{\xi} = [1, \boldsymbol{\mu}]' \tag{15}$$

の拡張平均ベクトルである．また，$\boldsymbol{H}$ は次式のように表すことができる．

$$\boldsymbol{H} = [\boldsymbol{b}, \boldsymbol{A}]' \tag{16}$$

ここで，$\boldsymbol{A}$ は $n \times n$ の変換行列，$\boldsymbol{b}$ はバイアスである．この変換行列 $\boldsymbol{H}$ を最尤推定を用いて推定する．分布数 R のガウス分布 $\{m_1, m_2, \ldots, m_R\}$ で状態共有している $\boldsymbol{H}_m$ の推定は，一般化補助関数の定式化より平均の変換と共有したガウス分布から以下のように表される．

$$\sum_{t=1}^{T}\sum_{r=1}^{R} L_{m_r}(t)\boldsymbol{\Sigma}_{m_r}^{-1}\boldsymbol{o}(t)\boldsymbol{\xi}_{m_r}^{T} = \sum_{t=1}^{T}\sum_{r=1}^{R} L_{m_r}(t)\boldsymbol{\Sigma}_{m_r}^{-1}\boldsymbol{H}_m\boldsymbol{\xi}_{m_r}\boldsymbol{\xi}_{m_r}^{T} \tag{17}$$

そして，$q_{m_r}(t)$ は時刻 t のガウス分布 m_r を指し，$\boldsymbol{O}_T = \{\boldsymbol{o}_1, \ldots, \boldsymbol{o}_T\}$ は適応データを示している．

$\boldsymbol{H}_m$ を解くために，以下の 2 つの項目を定義する．

表 3　音響分析条件 (音声合成用)

sampling rate	16 kHz
frame length	25 ms
frame shift	10 ms
window	Blackman
feature vector	0–24 mel-cepstral , Δ mel-cepstral, Δ^2 mel-cepstral (total 75)

（1） 式 (17) の左辺は変換行列から独立で $\boldsymbol{Z}$ は，

$$\boldsymbol{Z} = \sum_{t=1}^{T}\sum_{r=1}^{R} L_{m_r}(t)\boldsymbol{\Sigma}_{m_r}^{-1}\boldsymbol{o}(t)\boldsymbol{\xi}_{m_r}^{T} \tag{18}$$

とする．

（2） 変数 $\boldsymbol{G}^{(i)}$ は，以下の要素で定義される．

$$g_{jq}^{(i)} = \sum_{r=1}^{R} L_{m_r}(t)\boldsymbol{\Sigma}_{m_r}^{-1}\boldsymbol{o}(t)\boldsymbol{\xi}_{m_r}^{T} \tag{19}$$

ここで，

$$\boldsymbol{V}^{(r)} = \sum_{t=1}^{T} L_{m_r}(t)\boldsymbol{\Sigma}_{m_r}^{-1} \tag{20}$$

そして，

$$\boldsymbol{D}^{(r)} = \boldsymbol{\xi}_{m_r}\boldsymbol{\xi}_{m_r}^{T} \tag{21}$$

以上の 2 つ定義から $\boldsymbol{H}_m$ は，

$$\boldsymbol{h}_i^{T} = \boldsymbol{G}_i^{-1}\boldsymbol{z}_i^{T} \tag{22}$$

のように計算することができる．ここで，$\boldsymbol{h}_i$ は $\boldsymbol{H}_m$ の i 番目のベクトルで，$\boldsymbol{z}_i$ は $\boldsymbol{Z}$ の i 番目のベクトルである．推定された変換行列 $\boldsymbol{H}$ を複数の分布で共有することにより，適応データのない分布に対しても適応後の平均ベクトルが得られ，モデル全体を適応することができる．なお，適応データが少ない場合に，$\boldsymbol{A}$ として対角行列を用いる手法も提案されており次節の実験で比較を行う。

4.2 話者適応実験

提案手法に MLLR を適用した場合に，適応に必要な音声の量および適応による音声品質の向上度合いを確認するために客観評価尺度 PESQ [9] による評価実験を行った。PESQ 値は – 0.5 〜 4.5 までの値であり，値が大きい程良質な音声である．

4.2.1 実 験 条 件

音声合成用の不特定話者音響モデルは，男性 54 名，女性話者 54 名が発声した音素バランス文 5400 発声を用いて学習した状態共有 triphone HMM (2000 状態 1 混合) である。特定話者モデルとしては FAK が発声した音素バランス文 453 発声を用いて学習した状態共有 triphone HMM

(800 状態 1 混合) を使用した。また，適応データとして，FAK が発声した音素バランス文を用いた．表 3 に音響分析条件をを示す．評価には，FAK が発声した新聞記事読み上げ文 21 発声を用いた．なお，不特定話者モデルの学習データ，適応データはそれぞれ異なる発声内容である．

4.2.2 実　　験

まず，適応データを用いて，MLLR 法により，不特定話者モデルを FAK に適応させた適応話者モデルから評価用の合成音声を作成した．このとき，式 13 の変換行列 $\boldsymbol{A}$ に関して，全成分を作成する場合と，対角成分のみを作成する場合の比較を行った。また，適応データ量は，3 ～ 120 (s) の発声を用意した．なお，ピッチ情報は原音声から抽出したものを用いた．

表 4　適応データ量と音声品質の関係

適応データ (秒)	3	5	10	20	60	90	120
全成分	1.71	1.99	2.13	2.16	2.16	**2.17**	2.15
対角成分のみ	**2.13**	2.09	2.11	2.10	2.11	2.10	2.11

表 5　特定話者モデル及び不特定話者モデルの音声品質

特定話者モデル	**2.15**	不特定話者モデル	**1.98**

4.2.3 結果および考察

実験結果を表 4 に示す。比較のため，特定話者モデル及び不特定話者モデルから作成した合成音声の品質も表 5 に示した。

表 4 より，対角成分のみの変換行列を用いた場合は，適応データ数に関係なくほぼ一定の音質となる．逆に、全成分の変換行列を用いた場合は，適応データ数の増加に伴い，音質も向上することがわかる．これより，適応データ数が少ない場合には対角成分のみの変換行列を，適応データ量が多い場合には全成分の変換行列を用いる方が良いことがわかる．また，対角成分のみの変換行列を用いた場合，3 秒の敵応データを用いることで，表 5 の特定話者モデルと同等の品質がえられていることがわかる．そこで，次章の実験では 3 秒の適応データを用い，対角成分のみの変換行列で適応した合成用音響モデルを用いることにする．

5. 評 価 実 験

提案する音声途切れ補間手法の有効性を示すために，主観評価実験および客観評価実験を行った。まず，途切れ区間のある音声 (以下，途切れ音声と表記) と提案手法を用いて音声途切れ区間を合成音声で補間した音声 (以下，補間音声) の聴き易さを比較するために，対比較試験による主観評価実験 [6], [7], [10] を行った。次に，途切れ音声と補間音声の品質を調べるために，PESQ [9] を用いた客観評価実験を行った。

表 6　主観評価実験結果 (CMOS)

	missing section (s)		
	0.5 ～ 0.7	1.0 ～ 1.2	1.5 ～ 1.7
recovered speech	1.2	0.7	1.0

表 7　客観評価実験結果 (PESQ score)

	missing section (s)		
	0.5 ～ 0.7	1.0 ～ 1.2	1.5 ～ 1.7
speech with missing sec.	2.7	2.7	2.9
synthetic speech	2.1	2.1	2.2
recovered speech	3.3	3.3	3.5

5.1 実 験 条 件

音声合成用の音響モデルには，4. 2 節で作成した，不特定話者モデルを FAK の 3 秒の音声で MLLR 法 (対角成分のみ) により話者適応した適応音響モデルを用いた。評価には，FAK が発声した新聞記事読み上げ文 21 発声を用いた。各発声に対し 0.2 (s) の音声途切れ区間を発生させ，途切れ音声を作成した。音声途切れ開始位置は 0.5 (s)，1.0 (s) ，1.5 (s) とし，各 7 発声である。なお，これらの発声は Missing Feature Theory を適用した音声認識により，正しく認識できている。

5.2 主観評価実験

途切れ音声と補間音声の聴き易さを調べるために，対比較試験による主観評価実験 [6], [7], [10] を行った。被験者は男性 14 名で，受聴にはスピーカーを用いた。始めに，途切れ音声を再生し，続いて補間音声を再生した。評価は途切れ音声に対する補間音声の品質で行った。この評価は，5 段階 (−2: much worse，−1: worse，0: about the same，1: better，2: much better) で評価し，CMOS (Comparison Mean Opinion Score) [10] 値を求めた。

評価結果を表 6 に示す。途切れ位置によりばらつきがあるものの，補間音声は全ての条件において，途切れ音声より良い評価が得られていることがわかる。しかしながら，0 より評価の低い音声も複数存在した。それらの音声を調査すると，(1) 途切れ区間がたまたま母音＋促音であったため途切れによる影響がない場合，(2) 合成音声の質が極端に低い場合，の 2 つに分類された。

5.3 客観評価実験

途切れ音声と補間音声の品質を調べるために，PESQ [9] 値を用いた客観評価を行った。原音に対する各音声の評価結果を表 7 に示す。比較のため 1 文全体を合成音により作成した音声の PESQ 値も示した。

表より，途切れ開始位置に関係なく音声が途切れると音声品質が激しく劣化することがわかる。また，合成音声の

品質はさらに悪いことがわかる。しかし，途切れ区間をその合成音声で補間することで，音声品質は改善されている。主観評価実験で 0 より評価の低かった音声を含め、補間により PESQ 値が低下する音声はなかった。

5.4 考　　察

両実験結果より，パケット損失や揺らぎにより，音声が途切れた場合，著しく音声品質を劣化させ，さらに，音声情報の一部を損失することから人間にとって大変聴きにくい音声になってしまうことがわかった。これに対し，提案手法により音声を補間することで主観評価および客観評価において高い改善率が得られた。一方で，主観評価においては，CMOS 値か 0 より低くなってしまった補間音声も存在した。この現象の主な原因は合成音声の品質が悪いことであり，その傾向は，客観評価実験での合成音声の PESQ 値からも見て取れる。現在，本実験で用いた合成音声が，音源信号としてインパルス系列を用いていることが原因のひとつと考え，残差信号を用いた合成音作成の検討を進めている。

6. む　す　び

本論文では，従来のパケットロス隠蔽法に比べより長い区間の音声途切れ及び複数音素の欠落に対応するために，統計的手法を用いた音声認識と音声合成を利用した音声途切れ補間手法を提案した。提案手法は，(1) Missing Feature Theory に基づく音声認識により，途切れ区間の音素片列を前後の言語情報及び音響情報から推定し，(2) 推定した音素片列に基づき HMM 音声合成を行い，途切れ区間及び前後の音声波形を生成し補間する。提案手法により途切れ区間を補間した音声を用い，評価実験を行ったところ，客観評価においては全ての音声において補間により音声品質が改善することが確認された。また，主観評価実験においても平均的には高い改善効果が得られた。

しかし，補間により品質が下がると判断される音声もいくつか存在した。この現象の主な原因は，合成音声の質の問題であり，合成音声の品質向上が今後の課題として残された。また，音声途切れ区間の推定に誤りがあった場合の評価実験も今後行っていく必要がある。

謝　　辞

本研究は，平成 15 年度工学部研究プロジェクトとして研究助成を賜りました．関係各位に深く感謝の意を表しお礼申し上げます．本研究の一部は文部科学省科学研究費，基盤研究 (B)(2)14350204，14380166，若手研究 (B)15700163，国際コミュニケーション基金，放送文化基金の補助も受けております。また本研究は，工学研究科博士前期課程の小林邦嘉君の献身的な努力によっています．ここに記して，感謝の意を表します．

文　　献

[1] ITU–T Recommendation G.711 – Appendix I： A high quality low-complexity algorithm for packet loss concealment with G．711． Sep，1999．

[2] 黒岩眞吾，加藤恒夫，清水徹，樋口宣男，音声信号の途切れ・オーバーフローへの Missing Feature Theory の適用．日本音響学会講演論文集，pp．149–150，1999．

[3] Endo, T., Kuroiwa, S.， Nakamura, S.，Missing Feature Theory applied to Robust Speech Recognition over IP Network．Proc．Eurospeech，Vol．4，pp．3081–3084，2003．

[4] 益子貴史，徳田恵一，小林隆夫，今井聖，動的特徴を用いた HMM に基づく音声合成．信学論，Vol．J79–D–II，No．12，pp．2184–2190，1996．

[5] Imai，S.，Cepstral analysis synthesis on the mel frequency scale．Proc．ICASSP，pp．93–96，1983．

[6] 河原達也，住吉貴志，李昇伸，坂野秀樹，武田一哉，三村正人，山田武志，西浦敬信，伊藤克亘，伊藤彰則，鹿野清宏：連続音声認識コンソーシアム 2001 年度版ソフトウェア概要．SLP–43，pp．13–18，2002．

[7] 鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄，音声認識システム．オーム社，2001．

[8] C.J. Leggetter and P.C. Woodland， “Maximum likelihood linear regression for speaker adaptation od continuous-density hidden Markov models”，Computer Speech and Language, vol.9, pp.171-185, 1995.

[9] ITU–T Recommendation P.862：Peceptual evaluation of speech quality (PESQ)，an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs．Feb，2001．

[10] Keagy，S.，Integrating Voice and Data Networks．Cisco Press，2000．