論　文

# 大規模コーパスを用いた音声合成システムXIMERA

河井　恒[†,††a)]　戸田　智基[†,†††]　山岸　順一[†,††††]　平井　俊男[†]
倪　晋富[†]　西澤　信行[†]　津崎　実[†,†††††]　徳田　恵一[†,††††††]

# XIMERA: A Concatenative Speech Synthesis System with Large Scale Corpora

Hisashi KAWAI[†,††a)], Tomoki TODA[†,†††], Junichi YAMAGISHI[†,††††], Toshio HIRAI[†], Jinfu NI[†], Nobuyuki NISHIZAWA[†], Minoru TSUZAKI[†,†††††], and Keiichi TOKUDA[†,††††††]

**あらまし**　本論文では，ATR音声言語コミュニケーション研究所が開発した新しい音声合成システムXIMERAについて述べる．XIMERAは，これまでATRで開発された音声合成システム$\nu$-Talk及びCHATRと同様，コーパスベース方式を採用している．XIMERAの特長は，(1) 大規模な音声コーパス（日本語男声110時間，日本語女声59時間，中国語女声20時間，それぞれ単一話者），(2) HMMを用いた韻律パラメータのモデル化及び生成，(3) 知覚実験に基づく素片選択コスト関数の最適化，である．XIMERAの性能を評価するため，市販の音声合成システム10製品と合成音声の自然性を比較したところ，XIMERAが他のシステムより優れていることが示された．

**キーワード**　テキスト音声合成システム，コーパスベース方式，大規模コーパス，HMMを用いた韻律生成，知覚実験

## 1. まえがき

コーパスベース方式は，現在のテキスト音声合成システム（Text-To-Speech system，以下，TTS）の主流となっている．特に波形素片接続方式は，現時点で実用的なTTSを実現するための最も有望な方式であり，商用，実験用を問わず自然性の高い音声の合成を目的とするTTSにおいて広く採用されている．

ATRは，コーパスベース音声合成技術の発展において先駆的な役割を果たしてきた．その過程で，$\nu$-Talk [1] 及びCHATR [2], [3] という二つのTTSを開発した．筆者らは，これらに続く第3のTTSとして新たにXIMERA（ギリシャ神話に登場する怪獣Chimeraと同じ発音）を開発した [4]．

XIMERAの基本な枠組みは，$\nu$-Talk，CHATRを含む他のコーパスベース方式のTTSと同じであるが，これらと異なる特長として，(1) 大規模な音声コーパスの使用（日本語男声110時間，日本語女声59時間，中国語女声20時間，いずれも単一話者），(2) HMM (Hidden Markov Model) を用いたスペクトル・韻律ターゲットの生成，(3) 知覚実験に基づいて最適化された素片選択コスト関数，という3点を挙げることができる．コーパスベースの枠組みの中でこれらを含む要素技術を最適化することにより，合成音声の自然性の大幅な向上を図り，到達可能な上限を探ることが開発のねらいである．

本論文では，XIMERAの概要を紹介するとともに，主な要素技術の詳細を述べる．更に，XIMERAによ

† ATR音声言語コミュニケーション研究所，京都府
Advanced Telecommunications Research Institute International, Spoken Language Communication Research Labs., 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto-fu, 619-0288 Japan
†† KDDI研究所，ふじみ野市
KDDI R&D Labs., Fujimino-shi, 356-8502 Japan
††† 奈良先端科学技術大学院大学，生駒市
Nara Institute of Science and Technology, Ikoma-shi, 630-0192 Japan
†††† 東京工業大学，東京都
Tokyo Institute of Technology, Tokyo, 152-8550 Japan
††††† 京都市立芸術大学，京都市
Kyoto City University of Arts, Kyoto-shi, 610-1197 Japan
†††††† 名古屋工業大学，名古屋市
Nagoya Institute of Technology, Nagoya-shi, 466-8555 Japan
a) E-mail: hisashi.kawai@atr.jp

る合成音声の自然性を評価した実験についても述べる．

本論文の構成は以下のとおりである．**2.** では，コーパスベース式音声合成技術の歴史を背景に ATR で過去に開発された二つの TTS について述べる．**3.** では，XIMERA の概要について説明する．**4.** から **8.** では，XIMERA で用いられている要素技術と音声コーパスについて説明する．**9.** では自然性評価実験について述べ，最後に **10.** で本論文をまとめる．

## 2. ATR における TTS の研究開発

匂坂らは 1988 年，当時としては全く新しい音声合成の枠組みとして，不定長の音声合成単位を用いる手法を提案した [5], [6]．従来手法では，CV や VCV（C：子音（Consonant），V：母音（Vowel））などの固定的な単位の波形素片をあらかじめ音声コーパスより抽出・蓄積しておき，それらのみを音声合成に使用していた．これに対して，匂坂らの手法では，合成時に大きな音声コーパス（文献 [5], [6] では 5240 単語）全体の中から波形素片の長さを限定せずに波形素片系列を抽出する．この研究は，コーパスベース音声合成の端緒となった．また岩橋らは，音声コーパスに含まれるすべての波形素片の中から動的計画法を用いて音響的な評価基準に従って最適な波形素片系列を選択するアルゴリズムを提案した [7]．これらの研究成果の集大成として，音声合成システム $\nu$–Talk [1] が開発された（“$\nu$” は Non-Uniform Unit（NUU）に由来する）．

$\nu$–Talk で導入された音声合成の枠組みは革新的であったが，課題として (1) 波形素片がケプストラムでパラメータ化されているために合成品質がボコーダ的になる，(2) 音響的評価尺度と知覚尺度の対応関係の検証が不十分である，という問題が残された．後に匂坂は，コーパスベース方式の音声合成研究において，音声コーパス開発・合成アルゴリズム・品質評価基準を統合的に研究することの重要性を強調している [8]．

匂坂 [5] にやや遅れて，NTT の広川は，約 3 時間の音声コーパスを用いた波形素片接続型音声合成方式を提案した [9]．そこでは，合成ターゲットと候補となる音素セグメントとの間で音声の基本周波数（以下，$F_0$），$F_0$ の傾き，音素時間長，パワーの差に基づいて定義される評価尺度（後にターゲットコストと呼ばれるものである）を計算し，これを文全体で最小化するアルゴリズムが導入された．また，当初素片接続時には $F_0$ 操作は行われていなかったが，広川は後に PSOLA 法（Pitch Synchronouos Overlap Add）[10] を導入して韻律パラメータの操作を行った [11]．広川の提案したこのシステムは，今日の波形素片接続型 TTS システムの原型といえる．

ATR では，$\nu$–Talk に続く TTS システムとして CHATR が開発された [2]．当初，CHATR は音声合成研究用のワークベンチとして開発されたため，様々な種類の特徴量生成モジュールや波形生成モジュールを自由に組み合わせて音声合成の実験を行えるように設計されていた．波形生成モジュールとしては，$\nu$–Talk の方式（ケプストラムボコーダ）と並んで単純な波形素片接続方式も選択可能であった．初期の CHATR では，これらのモジュールが同等に扱われていたが，後に CHATR は代表的な波形素片接続型 TTS システムの一つとして知られるようになった．CHATR は，後にエジンバラ大の Festival [12] 及び AT&T の Next-Gen [13] に直接的・間接的に影響を与えた．また，最近では感情音声合成用のプラットフォームとしても用いられている [14]．

一方，ATR におけるコーパスベース TTS の研究と同時期に ATR 以外の研究機関でも素片選択手法についての研究が広く行われた（例えば，[15], [16]）．

CHATR は，限られたドメインでは非常に自然な音声を合成できたが，ドメイン外の入力テキストに対しては，十分な自然性を有する合成音を生成することが難しい場合があった．CHATR の課題として残された問題は，(1) テキスト処理・韻律のモデル化が脆弱である，(2) コーパス規模が小さい，(3) 素片選択用コスト関数が音響尺度のみで定義されており，人間の知覚特性との対応関係が検証されていない，ことであった．

これらの課題を解決し，波形素片接続型音声合成システムの限界を見極めることを目的として，ATR では新しい TTS システム XIMERA の開発が開始された．

## 3. XIMERA の概要

XIMERA のブロックダイアグラムを図 1 に示す．他の多くの波形素片接続型 TTS システムと同様，XIMERA は (1) テキスト処理モジュール，(2) スペクトル・韻律ターゲット生成モジュール，(3) 素片選択モジュール，(4) 波形生成モジュールという四つのモジュールから成り立っている．

XIMERA の対象言語は日本語と中国語である．コーパスベース方式の枠組みは，原理的には対象言語に依存しないが，現実的には，ほとんどの構成要素を特定の言語に特化して開発またはチューニングする必要が

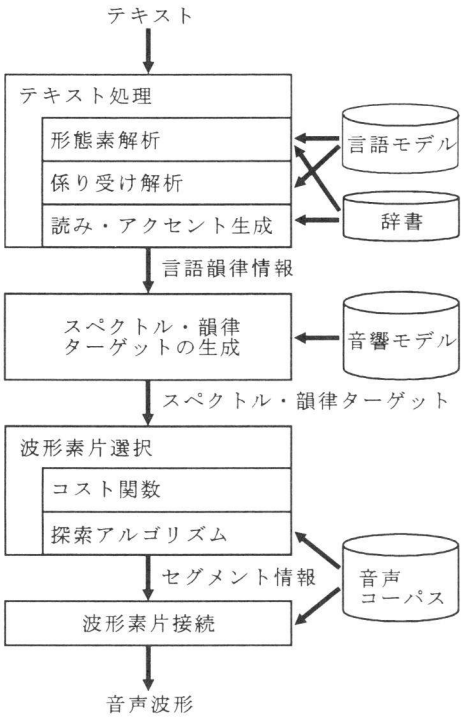



図 1 XIMERA のブロックダイアグラム
Fig. 1 Block diagram of XIMERA.

ある．言語依存となる構成要素は，具体的には，テキスト処理モジュール，スペクトル・韻律ターゲット生成用の音響モデル，音声コーパス，素片選択のためのコスト関数である．

人間の代わりとなり得る音声合成システムを実現するためには，多様な感情や発話様式の音声の合成が必要不可欠であるが，現時点での XIMERA は，ニュース読上げや感情表現を含まない人対機械の対話を主な適用分野としており，通常の読上げ発話スタイルでの合成に焦点を絞っている．

なお，XIMERA の構成は CHATR と基本的に同じであるが，ソフトウェアとしては独立にゼロから開発されたものであり，コードに共通性はない．

## 4. テキスト処理

日本語テキスト処理モジュールは，形態素解析，係り受け解析，読み・アクセント付与を行う三つのサブモジュールから成り立っている．

形態素解析は bigram 言語モデルをベースとしており，形態素辞書の規模は 24 万語，bigram 数は 25,000 である．係り受け解析サブモジュールでは，隣接単語間の接続強度を解析する．これらは $F_0$ パターン生成とポーズ位置の推定 [17] に利用される．読み付与処理では，同形語（複数の読みがある単語．例えば，「何」→「なに，なん」）の読みの決定，音便化処理，母音無声化処理を行う．アクセント付与処理では，アクセント句を構成する個々の形態素のアクセント型とアクセント結合規則に基づいて，アクセント句の境界及びアクセント型を決定する．

読み付与の性能を辞書及び bigram 作成に使用していない 5,000 文をテストセットとして評価したところ，モーラ精度 99.1%，アクセント句正解率 83.3%であった．ここで，モーラ精度は，100 × (正解数 − 挿入数)/全モーラ数，アクセント句正解率は，100 × 正解数/全アクセント句数とそれぞれ定義している．

中国語のテキスト処理は，テキスト正規化，形態素解析，ピン音付与を行う三つのサブモジュールから成り立っている [18]．テキスト正規化処理では，数字（例：電話，時刻）や単位・シンボル（例：cm，“$”，“–”）を適切なテキストに変換する．形態素解析は，日本語と同じく bigram 言語モデルにより行う．形態素辞書の規模は 59 万語，bigram 数は 3,200 である．ピン音付与処理では，形態素辞書に含まれる情報をもとに形態素からピン音への変換を行う．

ピン音付与の性能を辞書及び bigram 作成に使用していない 14,717 文をテストセットとして評価したところ，正解率は 97.0%であった．ただし，ここで用いたテストセットには，テキスト正規化が必要な文は含まれていない．

形態素解析のためのソフトウェアとして，日本語では『茶筌』[19] を，中国語では “MeCab” [20] をそれぞれ採用している．『茶筌』の標準の形態素辞書である “ipadic” には，アクセント型，アクセント結合属性などアクセント情報，無声化情報が含まれていないため，これらの情報は ATR で独自に付与した．また，人名，地名など固有名詞も増補している．日本語テキストの係り受け解析には，係り受け解析ソフトウェア『南瓜』[21] を使用している．

## 5. 音声コーパス

### 5.1 収録及び音声のデータベース化

日本語話者男女各 1 名，中国語話者女性 1 名による大規模な特定話者音声コーパスを収録した [22]．収録時間は，日本語男声が 110 時間，日本語女声が 59 時間，中国語女声が 20 時間である．時間長には発声前

表 1　日本語コーパス（男声）の内容
Table 1　Contents of the Japanese male speech corpus.

| Genre | Size in hours | Number of: utterances | phonemes | syllables |
|---|---|---|---|---|
| Sentence | 2.1 | 2,020 | 104,367 | 57,768 |
| News | 70.21 | 37,710 | 3,155,944 | 1,777,015 |
| Novel | 9.95 | 7,452 | 469,226 | 259,207 |
| Travel conv. | 12.11 | 20,761 | 573,060 | 323,341 |
| Word | 4.93 | 26,979 | 184,493 | 102,173 |
| Syllables | 0.34 | 2,493 | 8,086 | 5,753 |
| Voice check | 9.41 | 18,438 | 346,801 | 218,206 |
| Misc. | 1.63 | 2,106 | 77,703 | 43,628 |
| Total | 110.68 | 117,959 | 4,919,680 | 2,787,091 |

表 2　日本語コーパス（女声）の内容
Table 2　Contents of the Japanese female speech corpus.

| Genre | Size in hours | Number of: utterances | phonemes | syllables |
|---|---|---|---|---|
| Sentence | 2.42 | 2,012 | 104,157 | 57,734 |
| News | 20.46 | 12,350 | 872,204 | 488,491 |
| Novel | 17.27 | 12,979 | 753,758 | 413,366 |
| Travel conv. | 3.86 | 7,357 | 183,921 | 103,481 |
| Word | 5.71 | 26,868 | 183,364 | 100,819 |
| Voice check | 9.83 | 14,769 | 253,234 | 159,738 |
| Total | 59.55 | 76,335 | 2,350,638 | 1,323,629 |

表 3　中国語コーパス（女声）の内容
Table 3　Contents of the Chinese speech corpus.

| Genre | Size in hours | Number of: utterances | semi-syl. | syllables |
|---|---|---|---|---|
| News | 13.54 | 6,303 | 270,295 | 145,428 |
| Travel conv. | 6.65 | 8,407 | 153,581 | 83,600 |
| Misc | 0.07 | 424 | 644 | 1,748 |
| Total | 20.26 | 15,134 | 424,520 | 230,776 |

後の無音区間を含まない．日本語コーパスの発声内容は，ニュース，小説，旅行会話などであり，それぞれの発声数，時間長は表 1，表 2 に示すとおりである．一方，中国語コーパスの発声内容は，表 3 に示すとおり，新聞，旅行会話などである．

旅行会話の発声スタイルは，日本語では読上げスタイル，中国語では対話スタイルである．旅行会話以外は，両言語とも読上げスタイルである．発話者は，プロのナレータである．話者は，日本語では 40 名，中国語では 3 名の候補者の中から調音・韻律の正確さ，読み誤りの少なさなどを考慮して選択した．収録期間及び日数は，それぞれ日本語男声が 973 日及び 181 日，日本語女声が 307 日及び 95 日，中国語女声が 63 日及び 32 日である．

収録は防音室内で行った．マイクロホンは，高い SNR を確保するため，大口径ダイアフラム，単一指向性のコンデンサマイクロホン（Neumann TLM103）を使用した．収録音声は，サンプリング周波数 48 kHz，量子化精度 24 bit でディジタル化し，ハードディスク装置に直接記録した．

録音した音声は，発声ごとに分割した後，空調などに起因する低周波雑音を除去するためにカットオフ周波数 70 Hz の高域フィルタをかけ，3 dB のヘッドルームを確保して振幅を調整し，16 bit 整数に形式変換した上でファイルに格納した．

発声内容は，人手によって検査し，読み誤りを含む発声を排除するとともに，片仮名書き起こしの内容も同時に検査・修正した．

音素セグメンテーションは，基本的に環境非依存，特定話者 HMM を用いて自動処理により行った [23]．特定話者 HMM の学習データは，コーパスの一部を不特定話者 HMM によってセグメンテーションした結果を初期値とし，人手によって検査・修正することによって作成した．

筆者らの評価実験によると，自動セグメンテーションの精度は手動セグメンテーションとほぼ同程度である [23]．また，自動及び手動によりそれぞれセグメンテーションされた音声コーパスから合成音声を作成し，自然性を比較した実験によると，統計的には手動セグメンテーションの方が優れているが，実用的な差はわずかである [23]．ただし，自動セグメンテーションの精度は，話者，言語，収録状況など多くの要因により影響を受け，また，セグメンテーション精度が合成音声の品質に与える影響は，素片選択及び接続アルゴリズムによって大きく異なるため，筆者らの評価実験の結果は，必ずしも一般性があるとはいえないことに注意が必要である．

$F_0$ 抽出も基本的には自動処理であるが，韻律ターゲット生成用 HMM の学習データについては，人手による検査・修正を行った．

### 5.2　収録音声の品質評価

コーパス規模の拡大は，波形素片接続型音声合成の自然性を改善するための最も直接的な方法である．一方，大規模な音声コーパスの収録には数週間から数か月の期間を要するため，収録される音声には短期的・長期的な声質の変動が生じる．声質の異なる波形素片を接続すると不連続感を生じ，合成音声の音質劣化につながる．

音声収録装置の伝達関数の変動に対しては，長時間パワースペクトルに基づく周波数特性の等化フィルタが有効であることが知られている [24]．しかし，筆者

らの音声コーパスではこうしたフィルタの効果は限られていた[25], [26]. 考えられる理由の一つとしては，録音機器の設定は収録期間中一定に保たれており，唯一制御していなかった話者の口との距離変動は，聴感上の差異を生じるほどの周波数特性変動にはつながらないことが挙げられる．声質変動の原因としては，(1) 話者の体調によって生じる発声器官の変化，(2) 話者の精神状態によって生じる発声運動の変化，が可能性として挙げられる．

音質変動を測定するための音響的尺度に関して知覚実験に基づく検討を行った結果[27]によると，心理量である声質差スコアを最も精度良く予測できる音響的尺度の組合せは，$F_0$，発話速度，8〜16 kHz 帯域のパワー，MFCC（Mel-Frequency Cepstral Coefficent）距離，及び発声日差であった．

### 5.3 コーパス規模と合成品質との関係

コーパス規模と合成音声の自然性との関係は，素片選択のためのコスト関数の特性によって異なるが，XIMERA に関してこの関係を調べることは，TTS システム全般での傾向を予想する上で有意義であろう．ここでは，合成品質を物理量であるコスト値と心理量である MOS 値（MOS: Mean Opinion Score）の二つの側面から検討する．

図 2 は，日本語男声 110 時間コーパスを無作為に縮小することによってコーパス規模を変化させたときの，ATR 音素バランス文 503 文[28]（約 30300 音素）に対する平均コスト値の変化である．図中の Best 90%，Worst 10%とは，波形選択処理により選択された波形素片をコスト値でソートし，小さい方から個数ベースで 90%までをとったときの平均値，及び大きい方から 10%をとったときの平均値を表す．ここで，コスト値は **7.2** の式 (2) で定義されるものであり，平均は有声音のみを対象として計算した．

この図を見ると，Best 90%，Worst 10%ともコーパス規模 20 時間付近で平均コスト値がほぼ飽和していることが分かる．すなわち，これ以上コーパス規模を拡大しても，コスト関数が同じである限り音質の改善は期待できない．日本語女声コーパスに関しても同様の傾向であった．

一方，図 3 は，上記と同じ男声コーパスの規模を 7.5 分から 63 時間まで 10 段階に変化させて合成音声を作成し，知覚実験により自然性を評価した結果である[29]．図では，図 2 に合わせてコーパス規模 1 時間以上の実験結果のみを示している．前記の実験とは異

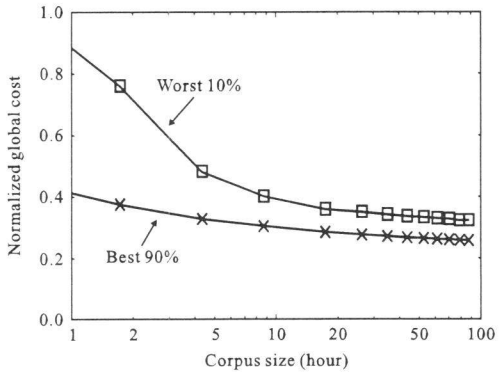



図 2 コーパス規模とコストとの関係
Fig. 2 Corpus size vs. cost. (The corpus is the Japanese male corpus. The test set is a set of 503 phonetically balanced sentences that is not included in the corpus.)

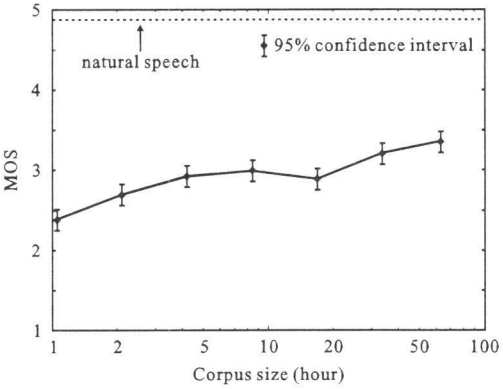



図 3 コーパス規模と MOS 値との関係
Fig. 3 Corpus size vs. MOS. (The corpus is the Japanese male corpus. The test set is a set of 8 sentences that is not included in the corpus.)

なり，縮小コーパスを作成する際に発声をランダムに並べ換える操作を行っていないため，コーパス規模が 1 段大きくなる際に新たなテキストジャンルが増える場合がある．刺激音声は ATR 音素バランス文 J セットから抽出した 8 文，評定者は日本人成人男女 17 名である．各評定者は，80 種類の合成音声を各 2 回及び 8 種類の自然音声を各 1 回，168 の刺激音声をヘッドホンを用いて両耳受聴し，自然性を 1：非常に悪い，2：悪い，3：普通，4：良い，5：非常に良い，の 5 段階で評価した．

図中のコーパス規模 17 時間の位置に見られるくぼみは，縮小コーパス作成手順の都合で生じた人為的な現象である可能性が高い．コーパス規模が 5 時間を超

表 4　HMM による韻律モデリングのためのトレーニングデータ
Table 4　Training data of HMM for prosody generation.

| Lang./Gender | Size in hours | Number of: utt. | labels* |
|---|---|---|---|
| 日本語 男声 | 3.26 | 1,809 | 157,778 |
| 日本語 女声 | 1.36 | 1,415 | 70,324 |
| 中国語 女声 | 2.45 | 1,680 | 19,928 |

*日本語：音素，中国語：声母，韻母．

えたあたりで，連続発声された素片系列の構成要素の一部が，素片選択時の予備選択によって捨てられる頻度が高くなったために不連続感が増大し始めたものの，コーパス規模 17 時間の付近で新たなテキストジャンルの音声が増え始めたために音素環境・韻律環境のカバー率が飽和から上昇に転じ，スペクトル・韻律ターゲットに近い素片が選択できるようになったためと考えられる．コーパス規模 17 時間付近のくぼみを無視すれば，MOS 値は 20 時間程度で飽和に近づいており，図 2 に示したコスト値に関する実験と矛盾のない結果となっている．

## 6. 韻律パラメータの生成

韻律パラメータ，すなわち，$F_0$，音素時間長，パワーは，HMM に基づく音声合成アルゴリズム [30]〜[32] によって生成される．日本語では，42 個の音素が各 5 状態からなる環境依存音素 HMM によってモデル化され，中国語では，60 の声母（initial，すなわち音節のはじめにある子音）及び韻母（final，すなわち音節から声母を除いた残りの部分全体）が各 5 状態の環境依存 HMM によってモデル化されている．生成された韻律パラメータは後段の素片選択モジュールに送られ，合成ターゲットとして使用される．HMM 学習用データの概要を表 4 に示す．

HMM の学習と韻律パラメータの生成には，名古屋工業大学と東京工業大学で開発された HTS [33]（HMM-based Triple S（Speech Synthesis System））を用いている．韻律パラメータ生成用音響モデルの学習には，HTK [34]，及び HTS の学習部を用いている．また，学習用データの音響分析のために SPTK [35] を用いている．

## 7. 素片選択

### 7.1 音声処理単位

XIMERA における最小の処理単位は，音素間だけでなく音素中央での接続を可能とするため，半音素を採用している [36], [37]．ただし，短い素片は一般に不連続感の増大につながりやすいことから，素片候補数が不足する可能性のある文頭・文末を除き，0.5 音素長の素片の使用を禁止している．また，日本語の場合は，発声単位が C-V 音節であり，知覚的にも子音–母音間の遷移部分の特徴が重要と考えられることから，C-V 境界での接続を禁止している．

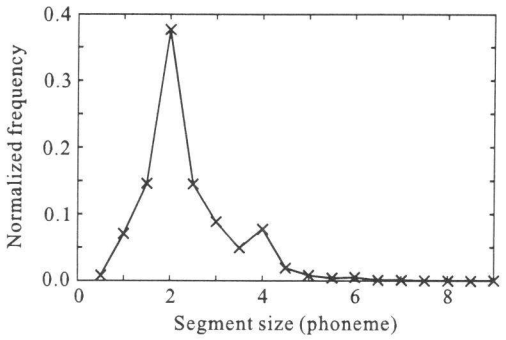



図 4　素片長の分布
Fig. 4　Distribution of segment size.

図 4 は，日本語男声コーパスの一部 88 時間分を素片データとして使用して ATR 音素バランス文 503 文の素片選択を行った場合の素片長の分布である．素片長の平均は 2.36 音素，標準偏差は 0.97 音素となっている．

中国語の場合，声母は子音，韻母は母音と同等として扱われる．四声情報も含めた声母，韻母の種類は，それぞれ 21,180 である．

### 7.2 コスト関数

1 文に対する素片選択のためのコスト関数は次のように定義される．

$$C_g = \left( \frac{1}{N} \sum_{i=1}^{N} \{C_t(t_i, u_i)\}^{p_t} \right)^{1/p_t} + \left( \frac{1}{N-1} \sum_{i=1}^{N-1} \{C_c(u_i, u_{i+1})\}^{p_c} \right)^{1/p_c} \quad (1)$$

ここで，$N$ は文中のターゲットの数，$C_t(t_i, u_i)$ はターゲットコスト，$C_c(u_i, u_{i+1})$ は接続コスト，$t_i$ と $u_i$ は $i$ 番目のターゲット及び選択された候補素片を，それぞれ表す．ただし，指数 $p_t$ と $p_c$ の値は，知覚実験 [38] の結果により 1.0 及び 1.5 としている．

ターゲットコストは次式で定義される．

$$C_t(t_i, u_i) = \sum_{j \in J_t} w_j \cdot C_j(t_i, u_i) \quad (2)$$

$$(J_t = \{F_0, \mathrm{dur}, \mathrm{cen}\})$$

ただし，$J_t$ の各要素は，それぞれターゲットと素片候補との $F_0$ 差，素片時間長差，スペクトル・セントロイド間のケプストラム距離を表す．式 (2) 及び (3) において，$w_j$ は対応するサブコストの重みである．

接続コストは次式によって定義される．

$$C_c(u_i, u_{i+1}) = \sum_{j \in J_c} w_j \cdot C_j(u_i, u_{i+1}) \qquad (3)$$

$$(J_c = \{F_0c, \mathrm{env}, \mathrm{spg}\})$$

ただし，$J_c$ の各要素は，それぞれ隣接素片候補間の $F_0$ 差，音素環境代替コスト，スペクトルの不連続性を示す．

音素環境の代替によるサブコスト値，音響的な尺度から各サブコストへのマッピング，及び各サブコストの重みは，知覚実験に基づいて最適化した [39]．その結果，コスト値と主観評価値の相関係数として，クローズ条件ながら $-0.84$ という高い値が得られている．

### 7.3 最適素片系列の探索

最適な音声波形素片の並びは，動的計画法により探索される [40]．コーパス規模が非常に大きい場合は，各素片候補の数が非常に多くなり（例えば，110 時間コーパスの場合，候補数は数万〜数十万となる），膨大な量の計算が必要になる．このため，ターゲットコストと接続コストの一部（音素環境代替コスト）に基づく予備選択を行い，計算量を削減している．

更に，対話処理のように応答時間，すなわちテキスト入力から音声出力開始までの遅延時間が短いことが必要とされる応用のために，文頭から文末に向かって一定時間長まで最適素片を探索した時点で，準最適な素片系列を文頭から順次出力するアルゴリズムの研究を行っている [41]．

## 8. 韻律変形と波形素片接続処理

音声コーパスから抽出した波形素片を単純に並べただけでは，韻律ターゲットと波形素片の $F_0$・音素時間長の誤差及び波形素片間の $F_0$ の不連続性のために，自然性劣化が生じる．しかしながら，こうした不自然性を解消するために信号処理的手法によって $F_0$ や音素時間長を変形すると，信号処理を行ったこと自体によって自然性劣化が生じる．

そこで，韻律誤差が残存することによる自然性劣化の程度と韻律変形による自然性劣化の程度を比較するために知覚実験を行った [29]．その結果，コーパス規模が 2 時間以上であれば，韻律変形による副作用が韻律誤差による自然性劣化を上回ることが分かった．XIMERA は 2 時間以上の大規模なコーパスを用いることを前提としているため，韻律変形を行わず，基本的に単純な音声波形の接続を行うこととした．ただし，音声波形の振幅の不連続性によって異音が生じるのを防ぐために，接続する素片境界の前後 5 ms の範囲で短時間相互相関係数が最大になる点において波形を接続している．

## 9. 自然性評価

XIMERA の合成音声の自然性を知覚実験によって評価した．比較対象として，2004 年 1 月時点で入手可能であった市販の TTS システム 10 製品を選定した．

各 TTS システムの話者は，女性に統一した．複数の女性話者が選択できるシステムでは，標準設定の話者を選択し，それがない場合は 1 名の話者を任意に選択した．XIMERA 及び CHATR のコーパス規模は，それぞれ 47 時間と 2.5 時間である．

音声合成の方式は，XIMERA と CHATR がコーパスベース方式（波形素片接続方式）であること以外は，コーパス規模も含めて基本的に不明であるが，図 5 中に I と示したシステムは，コーパスベースである旨カタログに記載されている．

テスト文は，10 ジャンルから各 8 文，合計 80 文を使用した．テキストのジャンルは，TTS の様々な応用場面を想定して選定したものであり，具体的には，旅行対話，音素バランス文，ニュース，コールセンター，構内放送，カーナビ・交通情報，天気予報，緊急時情報，娯楽情報，占い，である．

評価対象のすべての TTS システムにおいてテキスト処理の誤りによって読み・アクセントの誤りが見られた．これらの誤りは，事前に検査し，単語登録によって解消を図った．しかしながら，単語登録によっても間違いが解消されないもの，単語登録機能そのものが使用できないもの，単語登録によって別の箇所に問題が生じてしまったものについては，それ以上の手当は施さず，誤りを残したまま評価を行った．

評定者は健聴者 40 名で，全員が東京 23 区内で言語形成期を過ごし，かつ過去 3 年以上東京 23 区内で生活している者である．これは，標準語の韻律を正しく評定させるためである．性別は男女同数，年齢層は，20 歳台，30 歳台，40 歳台が 6 名または 7 名でほぼ均等とした．

評定者は，11 システム，10 ジャンル，8 文からなる合成音声刺激 880 個について 2 回ずつ，合計で 1760 回の試行を行った．各評定者は，合成刺激の自然性に関して，−3：とても悪い，−2：悪い，−1：どちらかというと悪い，0：どちらともいえない，+1：どちらかというと良い，+2：良い，+3：とても良い，の 7 段階の評定尺度に従って判断した．

刺激音声の提示順は，連続する 11 試行を 1 ブロックとした中に必ず一つの合成器による合成音声が一つずつ含まれる，という制約を除いてランダムとした．評定者には合成器が 11 種類あることについての情報は与えなかった．同じ文の異なる合成器による音声を連続して聞くことは，偶然の場合を除いて起こり得ないため，評定者が合成器，ジャンル，文の種類を予測することは不可能である．

実験の制御及び回答の収集は，ラップトップパソコン（IBM ThinkPad G40）を用いて行った．刺激音声は，パソコンのオーディオ出力からヘッドホン（SONY MDR-Z900）を通じて評定者の両耳に提示した．平均的提示音圧レベルは，70 dB（A 特性）とした．また，騒音環境を通常のオフィスに近づけるため，音圧レベル 48 dB（A 特性）の雑音をヘッドホンを通して実験中常に提示した．背景雑音は，電子協騒音データベース [42] No. 14 の計算機室のサンプルよりワークステーション 1 台による騒音を使用した．実験は，静かな会議室内で行った．

実験は八つのセッションに分割して実施した．各セッションの本試行数は 220 試行とし，11 の練習試行を冒頭に実施した．実験の進行ペースは，各評定者に任せた．半数が終了した時点で 2 分間の強制的な休憩を挿入したほか，評定者の判断で休憩を随時とることを許した．このため，1 セッションの所要時間は評定者ごとに異なり，おおよそ 1 時間から 1 時間 30 分程度であった．各評定者は 1 日 4 セッションを上限として，複数日にわたって実験に参加した．

各 TTS の評価値の平均値と標準偏差を図 5 に示す．コーパスベース方式であることが明らかなシステムについては，システムを表す記号に * を付したが，それ以外のシステムがコーパスベース方式であるかどうかは，不明である．図中で，棒グラフ上端から上下に伸びる線分の長さは，標準偏差の 2 倍に相当する．図から分かるように，XIMERA の自然性は他のシステムよりも優れている．また，Tukey-Kramer の HSD（Honestly Significant Difference）検定による合成器

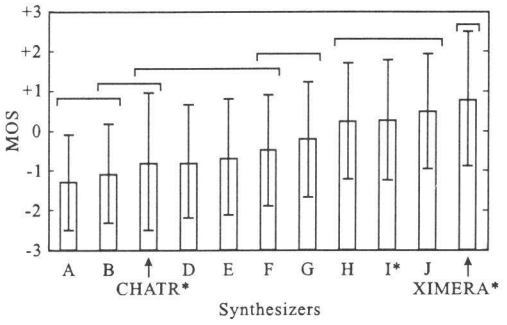



図 5　各種 TTS システムによる合成音声の自然性評価実験の結果（* はコーパスベース方式であることを示す．棒グラフ上の線分は，統計的に優位さのないグループを表す）

Fig. 5　The result of an evaluation experiment for naturalness between several synthesizers.

間の多重比較検定 [43] を行ったところ，XIMERA と次点の合成システム（図中 J）の間には有意な差が存在することが確認された（有意差水準 5%）．この多重比較検定の結果は図中に線分によって示してある．棒グラフの上に線分で囲われた合成器の間には，統計的な有意差が存在しないことを示す．

## 10. むすび

本論文では，ATR で開発された新しい波形素片接続型音声合成システム XIMERA について説明した．XIMERA の特徴とそれぞれに関連して得られた知見は以下のとおりである．

(1)　大規模音声コーパス

XIMERA では，合成音の自然性向上のため，従来のコーパスベース合成システムにはない大規模な音声コーパスを用いている．

最大 110 時間という大規模な音声コーパスを収録したことにより，コーパス規模を拡大すると録音時期差に起因する声質変化が生じ，それが新たな音質劣化の原因となることが分かった．声質時期差を測定するための音響尺度については，様々な検討を行ったが，主観尺度と高い相関を示す音響尺度はいまだ発見されていない．声質変動に対する現実的な対処方法としては，例えば，使用する音声コーパスのデータを刺激として知覚実験を行って発声セッション間の声質差のテーブルを作成し，これを素片選択に使用する方法が考えられる．

公式な評価実験は行ってはいないが，筆者らの印象によれば，声質変動の幅は話者によるところが大きい．

それゆえ，声質の経時変化の小ささは，話者選定時の重要な判断基準の一つと思われる．

(2) HMMを用いた韻律パラメータのモデル化及び生成

HMMベースの$F_0$制御においては，フレーズなど比較的長い単位での拘束が存在しないにもかかわらず，驚くほど自然性の高い$F_0$パターンが生成される．この手法の実用上の問題点は，トレーニングデータに含まれる不具合を除去し，クラスタリングのための質問を選定する作業に職人芸的技術が必要とされることである．こうした作業を効率化できるツール，あるいはこうした作業を全く必要としない完全な自動学習手法の開発がこの手法の今後の課題であろう．

(3) 知覚実験に基づく素片選択コスト関数の最適化

本システムで採用したコスト関数は，一般的な素片選択型音声合成で用いられるものと基本的に同じであるが，音素環境代替によるサブコスト値，音響的尺度からサブコストへのマッピング，サブコスト関数の重み等を知覚実験を積み重ねて決定したため，コスト値と主観評価値の相関係数$-0.84$（クローズ条件）という，人間の知覚にかなり近いものとなっている．

一方，コーパス規模対平均コスト値の実験によると，コーパス規模20時間付近でコスト値の飽和が観察された．このことは，コーパス規模を20時間以上に拡大しても合成音声の自然性向上にはつながらず，コーパス規模はこの程度で十分であることを意味する．同時にこのことは，現在のコスト関数・探索アルゴリズムの限界を示唆している．すなわち，飽和領域内で自然性に差があるにもかかわらず現在のコスト関数がそれを区別できていない可能性もある．あるいは，最適素片系列の探索アルゴリズムの限界のために探索空間の広がりが不十分である可能性もある．

(4) 音質の到達点

XIMERAの総合的な音質評価として，2004年1月時点で入手可能であった市販のTTSシステム10製品を比較対象として主観評価実験（MOS試験）を行った．その結果，XIMERAは他のいずれのTTSよりも優れていること，XIMERAと次点のTTSの間には，統計的に優位な差が存在すること，が確認された．

今後は，素片選択のためのコスト関数を改良すると同時に，合成音声に現れる不具合の分析とその改善を行っていく必要がある．

**謝辞** 本研究は独立行政法人情報通信研究機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである．

XIMERAへのHTSの組込み，及び高品質なHMMの作成に貢献して頂いた名古屋工業大学の吉村貴克博士（現在，株式会社豊田中央研究所），全炳河氏，University of Science and Technology of China (USTC)のYi-Jian Wu氏に感謝します．

## 文　献

[1] Y. Sagisaka, N. Kaiki, N. Iwahashi, and K. Mimura, "ATR ν-Talk speech synthesis system," Proc. ICSLP, pp.483–486, 1992.

[2] A.W. Black and P. Taylor, "CHATR: A generic speech synthesis system," Proc. COLING94, pp.983–986, 1994.

[3] N. Campbell, "CHATR: A high-definition speech re-Sequencing system," Proc. 3rd Joint Meeting of Acoustical Society of America and Acoustical Society of Japan, pp.1223–1228, Dec. 1996. Abstract: J. Acoust. Soc. Am., vol.100, no.4, Pt.2, p.2850, 5pSC14.

[4] H. Kawai, T. Toda, J. Ni, M. Tsuzaki, and K. Tokuda, "Ximera: A new TTS from ATR based on corpus-based technologies," Proc. of 5th ISCA Speech Synthesis Workshop, pp.179–184, June 2004.

[5] 匂坂芳典，"種々の音韻連接単位を用いた日本語音声合成，" 信学技報，SP87-136, March 1988.

[6] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," Proc. ICASSP, pp.679–682, 1988.

[7] N. Iwahashi, N. Kaiki, and Y. Sagisaka, "Speech segment selection for concatenative synthesis based on spectral distortion minimization," IEICE Trans. Fundamentals, vol.E76-A, no.11, pp.1942–1948, Nov. 1993.

[8] 匂坂芳典，"コーパス・ベース音声合成—音声科学知識に基づく合成システム構築技術の新パラダイム，" 音響講論集，pp.197–200, Sept. 1999.

[9] 広川智久，"波形辞書を用いた規則合成法，" 信学技報，SP88-9, May 1988.

[10] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun., vol.9, no.5–6, pp.453–467, Dec. 1990.

[11] T. Hirokawa and K. Hakoda, "Segment selection and pitch modification for high quality speech synthesis using waveform segments," Proc. ICSLP, pp.337–340, 1990.

[12] A. Black and P. Taylor, "Festival speech synthesis system: system documentation (1.1.1)," Technical Report HCRC/TR-83, Human Communication Research Centre, 1997. http://www.cstr.ed.ac.uk/projects/festival/

[13] M. Beutnagel, A. Conkie, J. Schroeter, Y. Stylianou, and A. Syrdal, "The AT&T next-Gen TTS sys-

tem," Proc. Joint Meeting of ASA, EAA, and DAGA, pp.18–24, March 1999.
[14] N. Campbell, "Towards synthesizing expressive speech: Designing and collecting expressive speech data," Proc. Eurospeech2003, pp.1637–1640, Sept. 2003.
[15] R.E. Donovan, Trainable Speech Synthesis, PhD. Thesis, Cambridge University, Engineering Department, 1996.
[16] A.P. Breen and P. Jackson, "Non-uniform unit selection and the similarity metric within BT's Laureate TTS system," Proc. Third ESCA/COCOSDA Workshop on Speech Synthesis, pp.201–206, 1998.
[17] 山岸順一，河井　恒，小林隆夫，"Naive markov model によるポーズ予測アルゴリズム，" 音響講論集，pp.349–350, Sept. 2005.
[18] J. Ni, H. Kawai, T. Toda, K. Tokuda, and N. Nishizawa, "A Chinese text-to-speech system at ATR," 音響講論集，pp.287–288, March 2005.
[19] 茶筌，2004. http://chasen.aist-nara.ac.jp/
[20] MeCab, 2004. http://cl.aist-nara.ac.jp/~taku-ku/software/mecab/
[21] 南瓜，2004. http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/
[22] 河井　恒，"音声合成用大規模音声コーパスの構築，" 信学技報，SP2005–19, May 2005.
[23] 河井　恒，戸田智基，"波形接続型音声合成のための自動音素セグメンテーションの評価，" 信学技報，SP2002–170, Jan. 2003.
[24] Y. Shi, E. Chang, H. Peng, and M. Chu, "Power spectral density based channel equalization of large speech database for concatenative TTS systems," Proc. ICSLP2002, pp.2369–2372, 2002.
[25] J. Ni, H. Kawai, and M. Tsuzaki, "Investigation of power spectral density based channel equalization," 信学技報，SP2003–67, 2003.
[26] J. Ni, H. Kawai, and M. Tsuzaki, "Detection and correction of the channel variability in a Mandarin speech corpus," Acoustical Science and Technology, vol.25, no.4, pp.303–306, 2004.
[27] H. Kawai and M. Tsuzaki, "A study on time-dependent voice quality variation in a large-scale single speaker speech corpus used for speech synthesis," Proc. IEEE 2002 Workshop on Speech Synthesis, Sept. 2002.
[28] 阿部匡伸，匂坂芳典，梅田哲夫，桑原尚夫，"研究用日本語音声データベース利用解説書（連続音声データベース），" Technical Report TR–I–0166，ATR 自動翻訳通信研究所，Sept. 1990.
[29] 戸田智基，河井　恒，津崎　実，"素片接続型テキスト音声合成における韻律変形の有効性，" 音響講論集，pp.201–202, Sept. 2003.
[30] K. Tokuda, T. Kobayashi, and S. Imai, "Speech parameter generation form HMM using dynamic features," Proc. ICASSP, pp.660–663, 1995.
[31] 徳田恵一，益子貴史，小林隆夫，今井　聖，"動的特徴を用いた HMM からの音声パラメータ生成アルゴリズム，" 音響誌，vol.53, no.3, pp.192–200, March 1997.
[32] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," Proc. ICASSP, vol.III, pp.1315–1318, May 2000.
[33] HTS, 2003. http://hts.ics.nitech.ac.jp/
[34] HTK, 2004. http://htk.eng.cam.ac.uk/
[35] SPTK, 2002. http://kt-lab.ics.nitech.ac.jp/~tokuda/SPTK/
[36] M. Beutnagel, A. Conkie, and A. Syrdal, "Diphone synthesis using unit selection," Proc. Third ESCA/COCOSDA Workshop on Speech Synthesis, pp.185–190, 1998.
[37] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano, "Unit selection algorithm for Japanese speech synthesis based on both phoneme unit and diphone unit," Proc. ICASSP, vol.I, pp.465–468, 2002.
[38] T. Toda, H. Kawai, M. Tsuzaki, and K. Shikano, "An evaluation of cost functions sensitively capturing local degradation of naturalness for segment selection in concatenative speech synthesis," Speech Commun., vol.48, no.1, pp.45–56, Jan. 2006.
[39] T. Toda, H. Kawai, and M. Tsuzaki, "Optimizing sub-cost functions for segment selection based on perceptual evaluations in concatenative speech synthesis," Proc. ICASSP, vol.I, pp.657–660, 2004.
[40] A.J. Hunt and A.W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," Proc. ICASSP, pp.369–372, 1996.
[41] 西澤信行，河井　恒，"短遅延音声合成のための素片選択法，" 信学技報，SP2004–48, 2004.
[42] 板橋秀一，"騒音データベースと日本語共通音声データ DAT 版，" 音響誌，vol.47, no.2, pp.951–953, Feb. 1991.
[43] J. Tukey, "The philosophy of multiple comparisons," Statistical Science, vol.6, pp.100–116, 1991.

（平成 18 年 4 月 3 日受付，7 月 6 日再受付）

**河井　恒**　（正員）

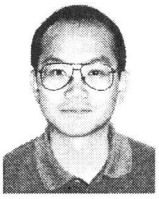
1984 東大・工・電気卒．1989 同大大学院博士課程了．同年国際電信電話（株）入社．2000 より ATR 音声言語コミュニケーション研究所へ出向．2003 音声合成研究室長．2004 KDDI 研究所へ帰任，音声処理グループリーダー．工博．音声合成，音声認識に関する研究開発に従事．2004 より IEEE 音声技術委員会委員．日本音響学会，IEEE 各会員．

**戸田　智基**　（正員）

1999 名大・工・電気卒．2003 奈良先端科学技術大学院大学博士課程了．同年，日本学術振興会特別研究員-PD（名工大）．2001～2003 ATR 音声言語コミュニケーション研究所研修研究員．2003 同研究所客員研究員．2003～2004 米カーネギーメロン大学客員研究員．2005 より奈良先端科学技術大学院大学情報科学研究科助手．工博．音声合成・分析・認識の研究に従事．2003 電気通信普及財団賞受賞．日本音響学会，IEEE，ISCA 各会員．

**山岸　順一**　（正員）

2002 東工大・工・情工卒．2006 同大大学院博士課程了．2004～2006 日本学術振興会特別研究員-DC1．2006 より日本学術振興会特別研究員-PD（東工大）．2003～2006 ATR 音声言語コミュニーケション研究所研修研究員．2006 より英エジンバラ大学客員研究員．工博．音声合成・認識，マルチモーダル・インタフェースの研究に従事．日本音響学会，IEEE，ISCA 各会員．



**平井　俊男**　（正員）

1988 阪大・工・原子力工卒．1990 同大大学院修士課程了．同年，（株）オージー情報システム総研（現オージス総研）に入社．1993～1997 ATR 音声翻訳通信研究所へ出向．その間，米ボストン大学客員研究員（1996）．1999（株）アルカディアに入社．2002 奈良先端科学技術大学院大学博士課程了．2004～2006 ATR 音声言語コミュニケーション研究所へ出向．博士（工学）．音声分析・合成に関する研究・開発に従事．日本音響学会会員．

**倪　　晋富**

1987 中国・ハルピン船舶工程學院・コンピュータ情報科学科卒．1990 同院修士課程了．同年，中國科學技術大助手，1992 同大学講師，1996 同大学助教授．2001 より ATR 音声言語コミュニケーション研究所客員研究員．博士（工学）（東大）．中国語音声合成，韻律モデリング，音声分析・合成の研究に従事．日本音響学会会員．

**西澤　信行**　（正員）

1998 東大・工・電気卒．2003 同大大学院・工・電子情報・博士課程了．同年 ATR 音声言語コミュニケーション研究所研究員．2006 より KDDI 研究所研究員．博士（工学）．音声合成に関する研究に従事．日本音響学会会員．

**津崎　　実**

1980 東大・文・第 IV 類卒．1982 同大大学院修士課程了．同年新潟大学人文学部助手．1985 東大文学部助手．1988 国際電気通信基礎技術研究所（ATR）に入社．以降 2004 まで，ATR 視聴覚機構研究所，ATR 人間情報通信研究所，ATR 音声言語翻訳研究所，ATR 音声言語コミュニケーション研究所研究員．1995～1996 英ケンブリッジ大学客員研究員．2004 より京都市立芸術大学音楽学部助教授．ATR 音声言語コミュニケーション研究所客員研究員兼務．聴覚に関する研究に従事．日本音響学会，アメリカ音響学会各会員．

**徳田　恵一**　（正員）

1984 名工大・工・電子卒．1989 東工大大学院博士課程修了．同年東工大電気電子工学科助手．1996 名工大知能情報システム学科助教授．2004 名工大大学院情報工学専攻教授．工博．音声言語情報処理，マルチモーダル情報処理，統計的学習理論の研究に従事．2001 電気通信普及財団賞，2001 本会論文賞，猪瀬賞各受賞．2000～2004 IEEE 音声技術委員会委員．日本音響学会，人工知能学会，情報処理学会，IEEE，ISCA 各会員．