発話速度と言語的特徴による変動を考慮した音素持続時間モデルを用いた音声認識

大河雄一[†] 伊藤彰則[††]
鈴木基之[††] 牧野正三[††]

本論文では，音声認識により生じる認識誤りのうち，持続時間が本来の長さと極端に異なるものを抑制する手段として，音素持続時間のモデルを用いる方法の検討を行った．そして，発話速度や言語的要因によってもたらされる持続時間の変動を考慮した，音素持続時間モデル化法と，その音声認識への適用手法の提案を行う．従来，音声合成の分野を中心として様々な音素持続時間の生成法が提案されているが，音声認識を目的として，発話速度の影響と音素の文中での位置や品詞などの言語的特徴の影響の双方を考慮に入れた音素持続時間のモデル化法や認識手法はなかった．本論文では，言語的特徴などを質問として用いた決定木により求められるクラスを単位とし，音素の持続時間と発話速度と相関の高い局所平均母音長の 2 次元正規分布として持続時間のモデル化を行うことで，様々な要因により変化する音素持続時間を高精度に推定を行う方法を提案する．また得られた持続時間の分布を，音声認識結果の N-best のリスコアリングに利用することで，認識率の改善が得られることを述べる．

A Phoneme Duration Model Considering Speaking-rate and Linguistic Features for Speech Recognition

Yuichi Ohkawa,[†] Akinori Ito,[††] Motoyuki Suzuki[††]
and Shozo Makino[††]

In this paper, we proposed a method of phoneme duration modeling for speech recognition. There was no usual method of duration modeling for speech recognition considering change by both speaking-rate and linguistic feature (phoneme location in sentence, part-of-speech et al.) Therefore, we modeled influence of speaking-rate by 2 dimension normal distribution of phoneme duration and local average of vowel duration. Each normal distribution is determined by tree based clustering with various question which include linguistic feature. We acquired 4.7% reduction of phoneme error rate by re-scoring of N-best hypothesis with proposed duration model.

1. はじめに

近年の音声認識技術の発展により，様々な分野で音声認識システムが用いられるようになっている．しかし，現在の技術水準においては，音声認識により 100%の認識精度が得られるまでには至っておらず，様々な要因により認識誤りが発生するため，これを減少させることが求められている．音声認識における誤りには様々な形態があるが，その一例としてモデルが局所的な音響的特徴にマッチして，誤って短い音素が挿入されたり，モデル化の精度が低いことにより音素のモデルが本来は正しくない区間にもマッチして，本来の音素が脱落してしまうような誤りが存在する（**図 1**）．一般にこのような認識誤りは，音素の持続時間が本来の長さに比べ極端に異なったものになる場合がある．またこれは，認識に用いられる HMM などの音響モデルが，きわめて限定的な持続時間制御機能しか持たないため，持続時間が極端に長かったり，短かったりするものであってもマッチしてしまうことも原因の 1 つといえる．そのため，音素の持続時間に関する知識を従来の認識法と合わせて利用することにより，こういった持続時間が本来の長さに比べて極端に異なる認識誤りを削減することが可能であると考えられる．

一方，音素の持続時間は様々な要因により変化するため，音声認識に用いる場合その高精度なモデル化が

† 東北大学大学院教育情報学研究部
Research Division, Graduate School of Educational Informatics, Tohoku University
†† 東北大学大学院工学研究科
Graduate School of Engineering, Tohoku University

要求される．たとえば発話速度は，直接的に音素の持続時間に影響を与えるものであるが，発話中においても大きく変化することが知られている．特に，近年新たな認識の対象として研究が行われている自然発話音声などは，発話速度の変動が大きい[1),2)]ため，発話速度の影響を高精度にモデル化する必要がある．また発話速度以外でも，音素の種類やコンテキスト，文中での位置や品詞といった言語的な特徴も持続時間に大きな影響を与えるため，持続時間のモデル化で考慮する必要があると考えられる．

従来，音声合成の分野では，合成音声の自然性確保を目的として持続時間のモデル化に関する様々な研究が行われている[3)~9)]．また，これらの知見を音声認識に用いる研究[10)~12)]も行われている．しかし，上であげた発話速度の影響と言語的な特徴の影響のすべてを考慮した持続時間のモデル化は行われていなかった．そこで本論文では，この音素の持続時間に影響を与える 2 つの特徴をともに考慮した持続時間モデル化の方法を提案する．また，このモデルを音声認識に用いる方法についても検討を行う．そのうえで，発話速度の変動が特に顕著だと思われる自然発話音声を対象に実験を行い，提案法の評価を行う．

なお，本論文の構成は次のとおりである．2 章では，音声認識に音素持続時間の知識を用いる方法の枠組みについて検討を行う．3 章では，音声認識を目的とした音素持続時間のモデル化の方法を提案する．4 章で，自然発話音声を対象に提案法の評価と考察を行い，最後に 5 章でまとめを行う．

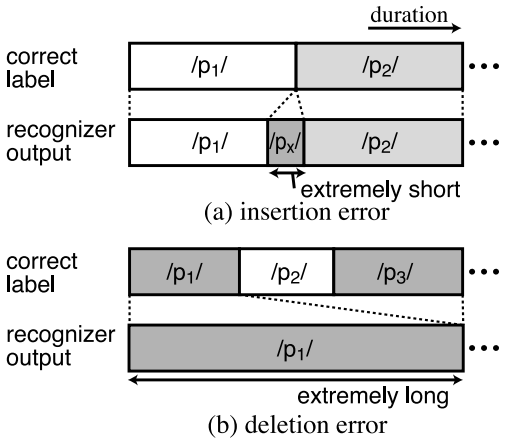

図 1 極端な持続時間をともなう認識誤りの例
Fig. 1 Example of recognition error with extreme duration.

2. 音素持続時間モデルを用いた音声認識の枠組み

従来，音声認識に音素の持続時間の情報を用いる方法としては，音響モデルに厳密な持続時間制御機能を与え，HMM の状態遷移確率を持続時間の関数としてガンマ分布や離散分布によりモデル化した方法[13),14)]が提案され，一定の効果が報告されている．

しかしこの方法では，音声認識のデコーディングの過程で持続時間のモデルを利用するという性質上，発話文全体の認識結果が得られるまでは確定できない平均モーラ長や平均母音長などで表される発話速度を考慮して，音素の持続時間を決定することはできない．また，音響モデルと一体になった持続時間のモデル化法であるため，発話文中での音素の位置など音素環境以外の要因による持続時間の変化を表現することも困難である．

そこで本論文では，**図 2** に示すように，音声認識によりあらかじめ認識結果の N-best 仮説を求め，リスコアリング[15)]を行うことにより持続時間モデルを利用することとした．この方法では，N-best 仮説から得られる音素系列を用い，発話の強制アライメントを行うことで，仮説ごとに音素の持続時間が得られる．また，あらかじめ文全体の音素持続時間を求めることができるため，発話速度の獲得も可能である．こうして得られた発話速度や言語的特徴を用いた高精度な持

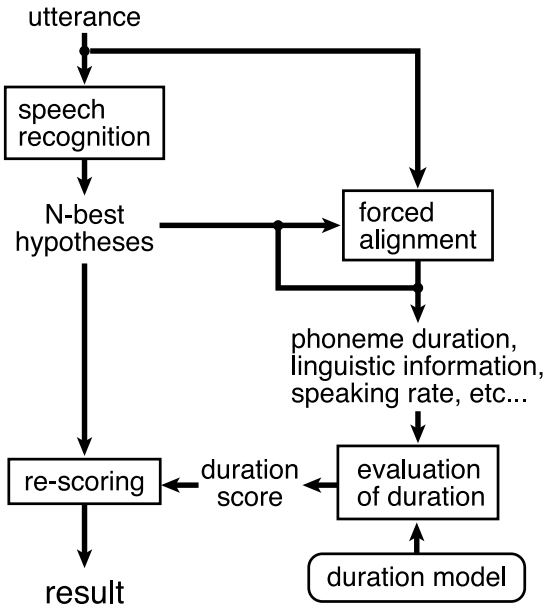

図 2 持続時間モデルを利用した音声認識の流れ
Fig. 2 A flow of speech recognition with phoneme duration model.

続時間モデルにより，各 N-best 仮説の音素持続時間の評価を行い，持続時間のスコアを出力する．この各 N-best 仮説のスコアを，認識時に得られている音響ゆう度や言語ゆう度とともに利用してリスコアリングを行い，N-best 仮説をソートしなおすことで，音素持続時間を考慮した自然発話音声認識が可能となる．本論文では，式 (1) により求まる最終的なスコア s_j を用いてリスコアリングを行い，最も高いスコアが得られるものを認識結果とした．

$$s_j = A_j + \alpha \cdot L_j + \beta \cdot W_j + \gamma \cdot D_j \qquad (1)$$

ここで，A_j，L_j はそれぞれ j 番目の N-best 仮説から得られた音響ゆう度，言語ゆう度であり，W_j は単語数である．また，α，β はこれらのゆう度の認識への寄与を調整する言語重み，および挿入ペナルティである．これらのパラメータは既存の音声認識システムでも一般に用いられる値である．一方，D_j は本論文で提案する持続時間スコアであり，γ はその寄与を決定する重みである．

3. 発話速度と言語的特徴を考慮した持続時間のモデル化

3.1 従来法の問題点

先に述べたように，音声合成の分野では合成音声の自然性を確保するための手段として持続時間推定法に関する多くの研究が行われている．また，音声合成で得られた知見を音声認識の高精度化のために利用しようとした研究もいくつか存在する．たとえば，線形近似による持続時間推定法[3] を利用した方法として，予備認識結果中の平均母音長を用いて音素持続時間の推定を行った研究[10] や，決定木による持続時間推定法[4] を応用し，決定木のリーフごとに音素の平均持続時間と分散を求め，認識結果を平均値からの偏りにより評価を行おうとした研究[11] などがあげられる．

しかしこれらの従来研究は，文中での発話速度の変動と言語的特徴の 2 つが音素の持続時間推定へ与える影響を考慮したものとなっていない．たとえば，線形近似による持続時間推定法を認識に利用した方法は，発話速度の影響は考慮しているものの，前章で述べた持続時間制御型の HMM と同様に音素単位でのモデル化法であり，発話文中での音素の位置や品詞などの言語的特徴は考慮されていない．また，音素持続時間の推定値は発話速度の 1 次関数により決定的に得られ，誤差の分布が考慮されていないため，認識結果から得られた音素の持続時間の評価を正しく行えない可能性がある．一方，決定木による持続時間推定法を認識に用いた方法では，決定木のリーフごとに求められる平均持続時間とその標準偏差を用い，入力音声の長さに合うよう標準偏差に一定の定数を乗じた値を平均持続時間に加え各音素の持続時間が決定される．そしてこの定数と，実際に得られた各音素の持続時間を得るために必要な標準偏差の係数との差により認識結果の評価を行う．その結果，この方法では，発話速度の変化が持続時間に与える影響と持続時間推定の誤差評価を標準偏差という 1 つのパラメータで行っていることになる．しかし，持続時間が発話速度から受ける影響は，誤差とは独立であり，音素の種類などによって異なるため，発話速度を考慮した持続時間モデル化としては良い方法とはいえない．また，文中での発話速度の変化に対応することもできない．

一方，発話速度と言語的特徴の双方をモデル化するため，発話速度や単語境界内での位置をいくつかのクラスに分け，持続時間の分布を求め認識に利用した研究[12] も存在する．しかし，この方法では，過学習が発生しないように，学習データの少ないクラスはモノフォンやバイフォンの分布が用いられることとなっており，学習データ量に比べ詳細にクラスを分割すると，各クラスの特徴を認識に用いることができない．そのため，学習データの量によりクラスが詳細になりすぎないようヒューリスティックに，分け方を変える必要があり，単語境界内での位置以外にも数多く存在する音素の持続時間に影響を与える言語的特徴を複数考慮したモデルを得ることは困難と考えられる．また，音素の持続時間は文中での発話速度と高い相関があると予測されるが，数個のクラスに分ける方法では発話速度の変化にともなう持続時間の連続的な変化をモデル化することはできない．

本論文では，これらの問題点に対し，文中で変化する発話速度を，局所的な平均母音長と音素持続時間の 2 次元正規分布としてモデル化を行い，言語的特徴を考慮した様々な設問を与え木構造クラスタリングを行うことによりモデルの詳細化を行う．これにより，発話速度と言語的特徴の双方の影響を考慮した音素持続時間のモデル化が可能と考えられる．

3.2 発話速度の影響のモデル化

一般に発話速度としては，単位時間あたりのモーラ数が用いられることが多い．しかし，モーラの単位に影響を及ぼす長母音・短母音の判別や促音の認識は比較的認識誤りが多いため，認識結果から求めるパラメータとしては安定性に欠ける．そのため本論文では，平均モーラ長との相関が比較的高い母音音素の平均持続時間[10] を発話速度を表すパラメータとして用いた．この際，文全体での平均を行うのではなく，式 (2) に

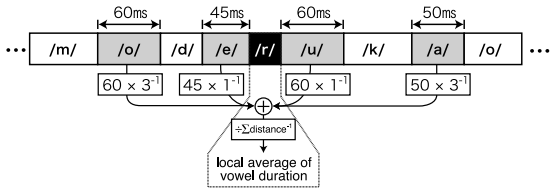

図 3 局所平均母音長の計算法
Fig. 3 A method to acquire local average of vowel duration.

示すように，当該音素と母音との位置により重みを与え，文中での発話速度の変動を考慮した，局所平均母音長を用いた（**図 3**）．

$$\hat{v}_i = \frac{\sum_{j \in vowel, i \neq j} w(i,j) \cdot d_j}{\sum_{j \in vowel, i \neq j} w(i,j)} \tag{2}$$

$$w(i,j) = \begin{cases} |i-j|^{-1} & \text{if } i \cdots j \text{ 間の短母音} \\ & \text{音素の数が } N_v \text{ 以下} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

ここで，d_i は文の先頭から i 番目に位置する音素の持続時間であり，$\hat{v}_i$ はその位置での局所平均母音長である．なお本論文では，局所平均母音長を求めるために短母音長のみを利用している．また，計算は文中のすべての母音音素を用いるのではなく，当該音素の周囲に限定し，各音素の前後各 N_v 個のみ用いた．本論文では，予備実験の結果などから前後 2 母音音素とした．さらに，中心音素が短母音音素であった場合にも，その母音音素の持続時間は評価対象であるため，局所平均母音長の計算には利用していない．

本論文では，発話速度による影響と他の要因による誤差を分けてモデル化するため，式 (4) に示すように音素持続時間 d と局所平均母音長 v の 2 次元正規分布としてモデル化を行った．

$$P(d,v) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}\boldsymbol{\zeta}^T \Sigma^{-1} \boldsymbol{\zeta}\right) \tag{4}$$

$$\boldsymbol{\zeta} = \begin{pmatrix} d - \mu_d \\ v - \mu_v \end{pmatrix} \tag{5}$$

ここで μ_d，μ_v は，音素持続時間と局所平均母音長の平均であり，Σ は共分散行列である．

また本論文では，認識結果を評価する目的でこのモデルを用いるため，式 (6) に示すように平均母音長について事後確率化を行った．これは認識結果から得られた量である局所平均母音長自体は，評価の対象ではないためである．

$$P(d \mid v) = \frac{P(d,v)}{\int_{-\infty}^{\infty} P(\theta, v) d\theta} \tag{6}$$

3.3 言語的特徴のモデル化

本論文では，音素の文中での位置や品詞などの言語的特徴により異なる持続時間分布をモデル化する手法として，木構造クラスタリングを用いた．そして，リーフに割り当てられた持続時間分布が最小となるように，あらかじめ用意した質問に従い分割を行い，持続時間分布選択のための決定木を作成した．アルゴリズムを以下に示す．

(1) 全学習用サンプルを含むリーフを 1 つ作成．
(2) 音素持続時間と局所平均母音長の 2 次元正規分布を求める．
(3) 開発用サンプルを与え，各サンプルごとに得られる対数ゆう度の総和を求める．
(4) 分割可能なリーフ l を選択する．
(5) N 個の質問 $Q_1 \ldots Q_N$ に対し (a)〜(c) を実行し，各設問に対する対数ゆう度の総和 $S_1 \ldots S_N$ を求める．
  (a) 選択された質問 Q_i によりリーフに割り当てられた音素サンプルを l_{yes}，l_{no} の 2 つに分割．
  (b) l_{yes}，l_{no} それぞれについて，音素持続時間と局所平均母音長の 2 次元正規分布を求める．
  (c) 開発用サンプルのうち決定木と質問 Q_i により l_{yes}，l_{no} に割り当てられるサンプルを与え，対数ゆう度の総和 S_i を求める．ただし，l_{yes} もしくは l_{no} に分割された学習用音素サンプルの数が最低分割サンプル数 N_{min} よりも小さい場合，または，各リーフの分布のいずれかの分散が 0 となる場合，質問 Q_i では分割不能とし $S_i = -\infty$ とする．
(6) 対数ゆう度の総和が最大となった設問を採用し，

$$\hat{i} = \arg\max_i S_i \tag{7}$$

設問 $Q_{\hat{i}}$ に従い学習サンプルの分割を行う．ただし，分割前のリーフ l の分布に対して求めた対数ゆう度の総和に比べ，$S_{\hat{i}}$ の改善が得られない場合，そのリーフは分割を行わず，分割不能とする．
(7) 分割可能なリーフが残っていない場合は終了，そうでなければ (4) に移動し分割を続ける．

以上により，**図 4** に示すような持続時間分布決定の

表 1 クラスタリングに使用した質問
Table 1 Questions for classification.

項目	詳細	種類
文平均の母音長 局所平均の短母音長	30 ms 以下，40 ms 以下，50 ms 以下，60 ms 以下， 70 ms 以下，80 ms 以下，90 ms 以下，100 ms 以下	各 8
音素の特徴， 前コンテキストの特徴， 後コンテキストの特徴	母音，長母音，有声子音，無声子音，半母音，はつ音，促音， 破裂音，摩擦音，破擦音，鼻音，弾音，接近音，硬口がい， 軟口がい，歯茎，両唇，声門，無音，よう音	各 20
音素の種類	/a/，/i/，/u/，/e/，/o/，/a:/，/i:/，/u:/，/e:/，/o:/， /w/，/y/，/r/，/p/，/t/，/k/，/b/，/d/，/g/，/m/，/n/， /h/，/f/，/ts/，/ch/，/s/，/sh/，/z/，/j/，/N/，/q/， /ry/，/ky/，/by/，/gy/，/my/，/hy/，/py/，/ny/，/dy/	40
単語の品詞	名詞，代名詞，動詞，形容詞，助詞，助動詞，副詞，連体詞， 接続詞，感動詞，形状詞，記号，接頭辞，接尾辞， 言いよどみ	15
単語内での位置	先頭，先頭 2 音素以内，末尾，末尾 2 音素以内	4
文内での位置	文頭，文頭 2 音素以内，文頭 5 音素以内， 文頭 10 音素以内，無音直後，無音後 2 音素以内， 無音後 5 音素以内，無音後 10 音素以内	8
単語長	5 音素以下，7 音素以下，9 音素以下	3
合計		146

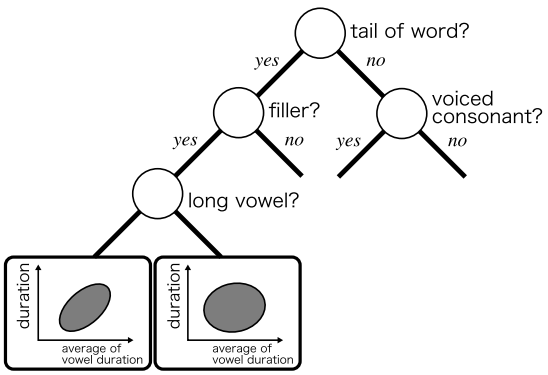

図 4 提案する持続時間モデル化法のイメージ
Fig. 4 An example of proposed method of duration modeling.

ための決定木を学習することが可能である．このアルゴリズムでは，クラスタリングによる過学習の問題を回避するため 2 つの分割停止条件を導入している．1 つは，開発用サンプルに対して，対数ゆう度の改善が得られない場合である．もう 1 つは，学習用サンプルから各リーフに割り当てられたサンプルの数が，事前に設定した定数 N_{min} を下回った場合である．この最低分割サンプル数は，予備実験により実験的に決定を行う．

なお本論文では，クラスタリングの過程で用いる質問 $Q_1 \ldots Q_N$ として，**表 1** に示す 146 種類の質問を用意した．

3.4 N-best 仮説の持続時間の評価

本論文ではここまで述べた持続時間のモデルをリスコアリングに用いるため，式 (8) に示す持続時間スコア D_j により，認識結果の N-best の各仮説を持続時間で評価した．

$$D_j = \frac{\sum_{i=1}^{M_j} \log P(p_{ij} \mid \hat{v}_{ij}, \mathcal{N}_{ij})}{M_j} \tag{8}$$

ここで，p_{ij}，$\hat{v}_{ij}$ は，音声認識から得られた N-best 仮説の j 番目の仮説に含まれる文頭から i 番目の音素の持続時間と，その位置での局所平均母音長であり，$\mathcal{N}_{ij}$ は，決定木により決定されたその音素サンプルの持続時間の分布である．また，M_j は N-best 仮説に含まれる音素の数である．

この式ではまず，決定木により選択された N-best 仮説の各音素の持続時間分布と局所平均母音長を用いて，各音素の持続時間の対数ゆう度を求める．そして，得られた対数ゆう度の文全体の和を音素数 M_j で割り正規化した値を持続時間スコアとした．単なる対数ゆう度の和ではなく平均値を用いたのは，対数ゆう度の和の大きさが，持続時間の優劣よりも音素数に強く依存するためである．

こうして求められた持続時間スコアを，前節の式 (1) でリスコアリングに用いた．このとき，音響ゆう度や言語ゆう度の値と大きさを合わせるため，持続時間スコアの寄与 γ には，全発話一定の持続時間重み w_d に N-best 仮説の平均音素数を乗じた値を与えた．

$$\gamma = w_d \cdot \frac{\sum_{k=1}^{N} M_k}{N} \tag{9}$$

4. 評価実験

4.1 持続時間推定性能の評価実験

前章で提案した持続時間分布モデル化法の性能を評価するため，音素持続時間推定実験を行った．

実験には日本語話し言葉コーパス（CSJ）[16] から男性による学会講演の音声を用いた．このコーパスは，文の分割位置がラベル付けされていないため，各講演を 0.5 秒以上の無音部ごとに分割し，分割された発話が 20 秒を超える場合には，20 秒を超えて最初に出現した 0.2 秒の無音部で再分割を行った．この結果のうち，発話時間が 2 秒を超えたものを，文の単位として使用した．実験では，その中から話者が重複しないように，持続時間モデルの学習用サンプルとして 5 万文（400 話者・462 万音素），木構造クラスタリングを行う際に使用する開発用サンプルには 1 万 3 千文（100 話者・118 万音素），オープン実験の評価用には 1,109 文（100 話者・9 万 6 千音素）を選択し使用した．正解値として持続時間モデルの学習と評価に用いる各サンプルの音素持続時間のラベル情報は，高精度な持続時間ラベルの付けられた ATR 音声データーベース[17] B セットの音素バランス文 5,030 発話（10 話者）により学習された 4 混合の monophone HMM を用い，各発話の強制アライメントを行うことで求めた．得られた音素境界の情報のうち，322 点について目視により修正を行い，自動ラベリングで得られた境界位置との差を求めた結果，平均 0.9 ms，標準偏差 15.4 ms の誤差であった．菊池ら[18] の報告によれば，手動によるラベリングでもラベラの違いにより約 8 ms の誤差が生じるとされており，十分な精度が得られていると考えられる．

従来法との比較として，式 (10) に示すように音素持続時間の予測値 $\hat{d}$ を平均母音長 μ_v の 1 次関数として推定した線形近似による推定法[10] を用い，各音素の持続時間を推定しラベル情報との誤差を求めた．

$$\hat{d} = a\mu_v + b \tag{10}$$

ここで a，b は，回帰係数である．この比較により，本論文で提案する，文中での発話速度の変動を 2 次元正規分布で記述することの妥当性とコンテキスト以外の言語的特徴を用いることの効果の検証が可能である．この際，予測に用いる回帰係数を全音素に対し 1 つのみ求めた音素非依存モデルと，学習用サンプルに出現するすべてのコンテキストの種類 12,581 種類ごとに求めたコンテキスト依存モデルの 2 種類を学習した．このモデルの学習には，提案法との公平のため開発用サンプルも学習サンプルとして合わせて使用した．

図 5 木構造クラスタリングで得られた決定木の一部
Fig. 5 A part of decision tree acquired by proposed method.

提案法では，決定木により評価サンプル中の各音素の持続時間分布が得られる．全音素の持続時間の和が評価サンプルの文の長さと同じになるという制約下で，この分布から得られるゆう度の和が最小となる各音素の持続時間を持続時間の推定値とした．

提案法・線形近似法ともに持続時間の推定を行うにあたり，平均母音長が必要となるが，本実験では既知のものとして扱い，ラベル情報より求めた値を与えた．予備実験の結果から，提案法の最低分割サンプル数 N_{min} は 300 とした．

以上の条件から，提案法では，木構造クラスタリングにより 7,525 のクラスに分割された 2 次元正規分布の持続時間モデルが得られた．**図 5** に得られた決定木の一部を示す．図に示されるように，クラスタリングでは単語の末尾，感動詞，有声子音といった言語的特徴のほか，コンテキストなどの設問により分割が進む傾向が強かった．一方，決定木の末端では平均母音長に関する設問が用いられる傾向が強かった．

持続時間推定実験の結果，各方法により得られた持続時間推定値と正解値との誤差の分布を**図 6**，**図 7** に示す．なお，クローズ実験は，全学習用サンプルを推定の対象に行った実験であり，オープン実験は評価用

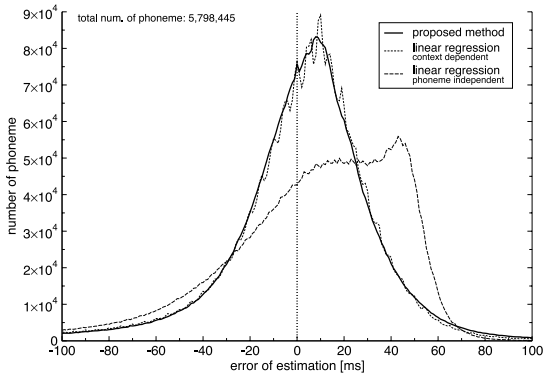

図 6 持続時間推定実験の結果（クローズ実験）
Fig. 6 Results of estimation experiments of phoneme duration (close test).

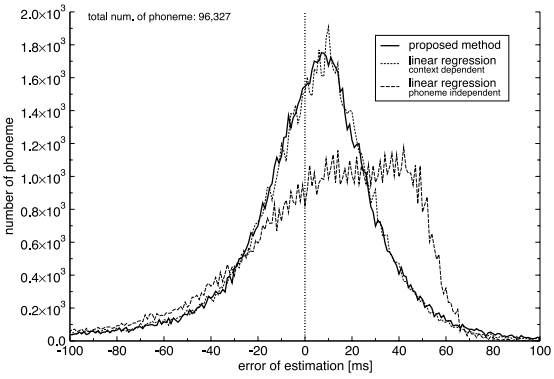

図 7 持続時間推定実験の結果（オープン実験）
Fig. 7 Results of estimation experiments of phoneme duration (open test).

サンプルの持続時間を推定した結果である．分布図の横軸は各音素の持続時間推定値がラベル情報の正解値からどの程度異なるかを表し，正の値は各推定法で得られた音素の持続時間がラベルの正解値より長い場合，負の値は正解値より短く推定された場合を表す．また，縦軸は誤差 1 ms ごとの音素の数である．

グラフより，提案法とコンテキスト依存の線形近似法は，オープン実験クローズ実験ともに近い分布をとっている．また，誤差の分布も正規分布に近い形状をして，比較的誤差の小さなサンプルが多いことが分かる．一方，音素非依存の線形近似法では，誤差の大きなサンプルがかなり多いことが分かる．また，グラフ中の各持続時間予測法で誤差のピークが正の値へ偏っていることが分かる．これは，各モデルを学習する際に用いたラベルの持続時間が，15 ms 未満の値をとらないことによる影響と考えられる．

また，これらの誤差分布を標準偏差で評価した結果を**表 2** に示す．提案法では，推定された持続時間分布

表 2 持続時間推定実験の結果（標準偏差）
Table 2 Results of estimation experiments of phoneme duration (standard deviation).

持続時間予測モデル	close 実験	open 実験
提案法	44.41 ms	44.81 ms
線形近似（コンテキスト依存）	44.28 ms	49.16 ms
線形近似（音素非依存）	56.85 ms	56.62 ms

表 3 持続時間推定誤差 100 ms ごとの音素の数（オープン実験）
Table 3 Number of phoneme each 100 ms of estimation error of phoneme duration (open test).

誤差の絶対値 [ms]	線形近似（コンテキスト依存）	提案法
〜 100 未満	92,741	93,135
100 以上 〜 200 〃	2,766	2,562
200 〃 〜 300 〃	553	459
300 〃 〜 400 〃	129	105
400 〃 〜 500 〃	70	37
500 〃 〜 600 〃	21	10
600 〃 〜 700 〃	13	8
700 〃 〜 800 〃	12	7
800 〃 〜 900 〃	5	4
900 〃 〜 1,000 〃	4	2
1,000 〃 〜	13	7

を用いて文の長さに合わせ伸縮を行ったため，推定誤差の平均値は 0 であり，線形近似を用いた場合も誤差の平均は最大で約 2.1 ms と非常に小さな値であった．そのため，ここでは標準偏差により各方式を比較する．

クローズ実験の場合，提案法，コンテキスト依存の線形近似ともに，高い性能が得られ，全サンプルから 1 組しか回帰係数を求めなかった場合に比べ標準偏差が約 12.5 ms 改善した．しかし，オープン実験では，コンテキスト依存の線形近似法は約 7.5 ms の改善しか得られなかった．一方，提案法では約 11.8 ms の改善が得られ，オープン実験での性能劣化はほとんどなかった．これは，誤差が極端に大きな誤りが提案法に比べて線形近似法に多く発生していることによる影響である．図 7 のオープン実験のうち，提案法とコンテキスト依存の線形近似の 2 つの推定法により得られた推定誤差の絶対値を，グラフの範囲外を含めて 100 ms ごとの音素数を求めた結果を**表 3** に示す．先の実験結果のグラフで示した区間である誤差の絶対値 100 ms 未満では，音素の数は 0.4%程度の差しかないが，推定誤差の絶対値が 100 ms 以上のすべての範囲で，提案法に比べ線形近似の音素数が 8〜110%多いことが分かる．また，オープン実験の推定誤差の標準偏差を発話ごとに求めプロットし，提案法とコンテキスト依存の線形近似法の比較を行った結果を**図 8** に示す．ここで点線の補助線は，各推定法による誤差の標準偏差がもう一方の推定法に比べ 50%以上大きい範囲を示

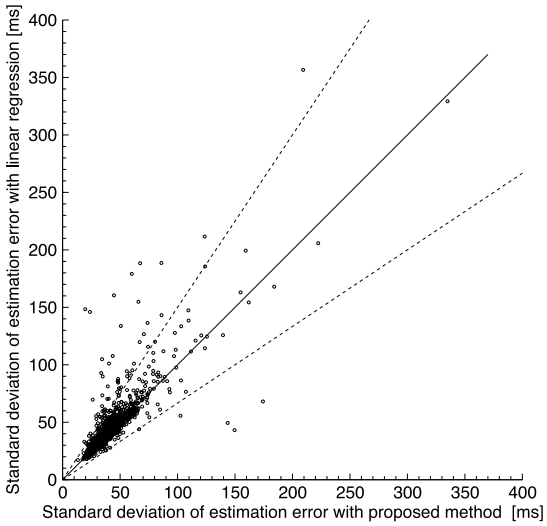

図 8 提案法とコンテキスト依存線形近似法の持続時間推定誤差の発話ごとの比較

Fig. 8 Comparison of estimation error of sentence between proposed method and linear regression (context dependent).

している．図から，提案法に比べ線形近似法による推定誤差の標準偏差が 50%以上大きい発話は 60 文であり，提案法が 50%以上となる発話が 5 文にとどまるのに対し大きく異なることが分かる．その原因として，学習用サンプル数に対しコンテキスト依存音素では持続時間推定の単位が詳細すぎて，過学習が発生していることが考えられる．

4.2 認識実験によるシステム全体の評価

提案法の認識実験での効果を見るため，自然発話音声の認識結果の 500-best の仮説に対し，持続時間スコアによるリスコアリングを行い，音素認識精度により評価を行った．

認識実験の対象は，前の実験と同様に日本語話し言葉コーパスの男性による学会講演音声を利用した．認識エンジンには Julius [19] バージョン 3.4.2 を利用した．提案法では言語的特徴に関する情報を必要とするため，約 1 万 6 千の語い数を持つ単語 3-gram を用い，品詞の情報などを出力した．この単語 3-gram は，持続時間モデル学習に用いた学習用・開発用のサンプルと同じ計 6 万 3 千発話（約 200 万形態素）から palmkit [20] を用い学習された．

音響モデルには，3 状態の PTM モデル[21] を用い，上記の持続時間モデル学習用サンプルのうち 6,261 文（300 話者）を用いて学習した．また，音響分析条件は**表 4** のとおりである．

以上の条件により，評価セット 1,109 文（100 話者）の認識を行い，各文 500-best の認識結果を得た．この N-best 仮説に対して，持続時間スコアの重み w_d を変えながらリスコアリングを行い，各重みで最も高いスコアの得られた仮説について音素認識精度を求めた．実験の結果を，**図 9** に示す．なお，木構造クラスタリングに 2 次元正規分布を組み合わせた提案法のほかに，比較対象としてコンテキスト依存および音素非依存の線形近似法により持続時間スコアを与えた結果も示す．ここで線形近似法では，持続時間の推定値が分布ではなく 1 次関数の解として決定的に得られるため，N-best 仮説内で観測された各音素の持続時間と推定値との差の絶対値の和を持続時間スコアとして与えた．また，その結果持続時間スコアのダイナミックレンジが提案法とは大きく異なるため，調整のため $\gamma = w_d/25$ とした．

表 4 音響分析条件
Table 4 Acoustic conditions.

特徴量	25 次元（12 MFCC + 12ΔMFCC + ΔPower）
窓関数	Hamming 窓
サンプリング周波数	16 kHz
フレーム幅	25 ms
フレーム周期	10 ms

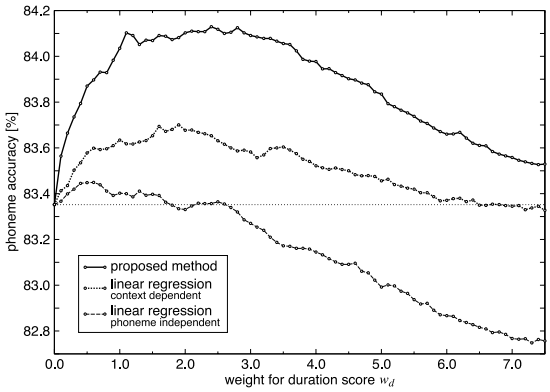

図 9 リスコアリングによる音素認識精度の変化

Fig. 9 Phoneme accuracy by weight of duration score.

この結果から，提案法では持続時間スコアの重みが 2.4 のとき，最も認識精度が向上し，持続時間の情報を用いない場合に比べて 4.7%の音素認識誤り削減率が得られた．一方で，コンテキスト依存の線形近似を用いた場合，重みが 1.9 のとき，最大の認識精度が得られるが，2.1%の音素認識誤り削減率にとどまる．音素非依存の線形近似では，ほとんど効果が得られないことが分かる．この結果について，母比率の差の検定を行うと，従来法であるコンテキスト依存の線形近似での認識精度の改善に比べ提案法での改善は 1%水準で有意に大きかった．

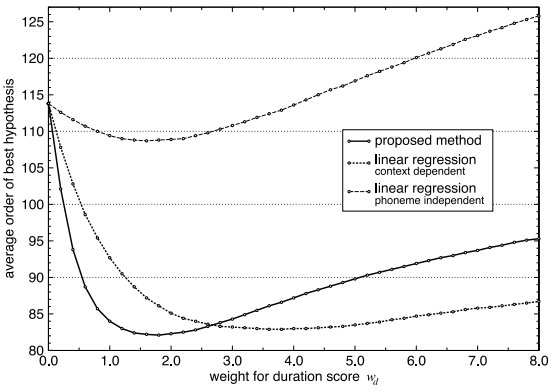

図 10 リスコアリングによる最良仮説の順位の変化
Fig. 10 Average order of best hypothesis by weight of duration score.

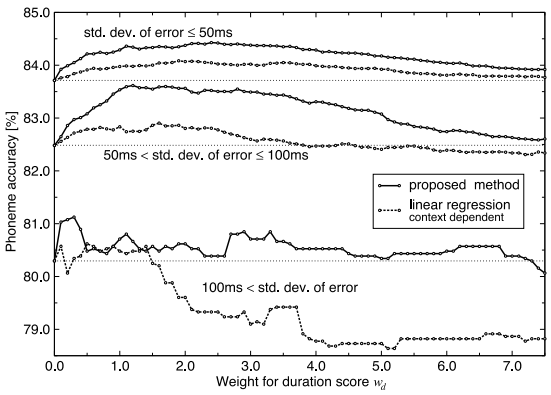

図 11 文ごとの持続時間推定誤差の標準偏差による提案法と従来法の効果の違い
Fig. 11 Difference of improvement by standard deviation of estimation error each sentence.

また同時に，各方式により N-best 仮説のうち最も正解に近い文がどの順位にあるのか変化を見た．結果を**図 10** に示す．仮説中での最も正解に近い仮説が N-best の何位までに現れるかを示す平均順位は提案法では，リスコアリング前の 113.8 位から最大約 82.1 位まで上昇した．一方で，コンテキスト依存の線形近似法でも，最大で約 82.9 位まで上昇しており，最良仮説の順位が必ずしも認識結果の向上につながっていないことが分かった．

4.3 考　　察

4.3.1 持続時間推定精度による効果の違い

持続時間推定の精度の違いが，リスコアリング後の認識精度にどういった影響を与えるのかを確認するため，4.2 節の認識実験の結果を文の持続時間誤差で分け，認識精度の評価を行った．図 8 の実験で求めた，文ごとの持続時間推定誤差の標準偏差を基準に，テストセットをコンテキスト依存の線形近似法で得られた推定誤差の標準偏差が 50 ms 以下の文，50 ms を超え 100 ms 以下の文，100 ms を超える文の 3 つのセットに分けた．各セットは，それぞれ 793 文，274 文，42 文である．その上で，提案法とコンテキスト依存の線形近似を用いて，重みを変えながら 3 つのセットのリスコアリングを行った．結果を**図 11** に示す．

グラフから，推定誤差の大きな標準偏差 100 ms を超えるセットに対するリスコアリングでは，提案法では大きくはないものの認識精度の一定の改善が得られるのに対し，線形近似法では認識精度の改善が得られる重みの幅が狭く，重み 1.5 以上では急速に認識精度の悪化が見られる．このことから，従来法に比べ提案法は，推定誤差の大きな発話にも比較的頑健であると考えられる．一方，推定誤差の比較的小さな 50 ms 以下と 50〜100 ms の 2 つのセットに対するリスコアリングでは，提案法，線形近似法ともに認識精度の改善が得られた．しかし，先に示した図 7 や図 8 の結果から，持続時間推定誤差の比較的小さな範囲では，提案法とコンテキスト依存の線形近似法に大きな推定能力差がないにもかかわらず，リスコアリング後の認識精度の改善に大きな差が出ていることが分かる．特に，50〜100 ms のセットで線形近似法では最大 2.4%の認識誤り削減率にとどまるのに対し，提案法では最大 6.5%の改善が得られている．このことは，推定誤差の分布を考慮しない従来法と異なり，提案法で 2 次元正規分布を用いて持続時間のモデル化を行った効果と考えられる．

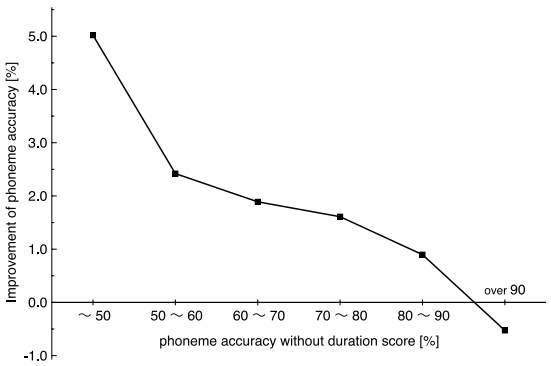

図 12 N-best 仮説の認識精度による提案法の効果の違い
Fig. 12 Difference of improvement by phoneme accuracy of N-best hypothesis.

4.3.2 N-best 仮説の認識精度による効果の違い

提案法により，どういった発話に対し効果があるのかを見るため，N-best 仮説の認識精度の違いにより，どの程度の認識率の改善が得られるのかの検討を行った．結果を**図 12** に示す．この図は，実験により最も良

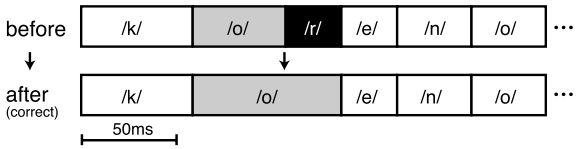

図 13 リスコアリングの前後で改善された例（1）
Fig. 13 An improved example with proposed method (1).

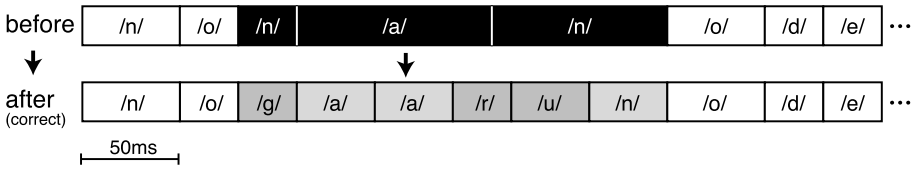

図 14 リスコアリングの前後で改善された例（2）
Fig. 14 An improved example with proposed method (2).

い性能の得られた重み 2.45 での音素認識精度が，持続時間スコアを利用しない場合の認識精度に比べどの程度改善されるのかを，認識精度ごとに見たものである．

結果から，音素認識精度が 50%以下の領域で最大の改善が得られ，5.0%の認識精度の改善が得られた．また，N-best 仮説の認識精度が上昇するに従い改善幅は小さくなり，認識精度が 90%を超える N-best 仮説に対して提案法を用いると逆に認識精度が低下することが分かった．このことから，提案法は認識精度が悪いほどその効果が高いことが明らかになった．

4.3.3 発話ごとの例

次に，提案法がどのようなサンプルに効果があるのか，認識率の改善があった個別のサンプルを調査した．**図 13**，**図 14** は，提案法により認識誤りが改善された例である．

図 13 は，ある文中で正しくは「声（こえ）の」と発声した部分である．それが持続時間スコアを用いなかった場合の結果は，「これの」という文が N-best の最初の仮説となっていた．この音声の持続時間を評価した結果，認識結果の/o/，/r/が正解文に比べ低いスコアとなり，正解文が得られた．

一方，図 14 は，「ものがあるのでは」という発話の一部である．認識結果では，2 つの/a/と/r/の一部が 1 つの/a/として，/r/，/u/，/n/が/n/として誤って認識され，持続時間が非常に長い音素となっていた．提案法によりこういった非常に長い音素が抑制され正解文が得られた．

5. ま と め

本論文では，音声認識を行う際に様々な要因で生じる認識誤りの中から，特に持続時間のおかしな誤りを排除するため，様々な要因で変化する音素の持続時間のモデル化を行った．従来，音声認識を対象として音素の持続時間をモデル化する際，文中での発話速度の変化や音素の出現位置や品詞などの言語的特徴による違いが正しく考慮されていなかった．

そこで本論文では，発話速度の影響を局所平均母音長と音素持続時間の 2 次元正規分布を用いてモデル化し，木構造クラスタリングにより言語的特徴を考慮する方法を提案した．また，このモデルを音声認識結果の N-best 仮説のリスコアリングに利用し，音声認識の高精度化を図った．

提案法を評価するため，日本語話し言葉コーパスの学会講演音声 1,109 文を対象とした音声認識の 500 best の仮説音素列に対し，持続時間モデルを用いて持続時間スコアを与え，リスコアリングに用いることで，音素認識精度を求めた．その結果，従来法であるコンテキスト依存の線形近似法により持続時間スコアを与えた場合では，持続時間の情報を用いない場合と比べて最大 2.1%の音素認識誤り削減率が得られるにとどまるのに対し，提案法では，最大 4.7%の音素認識誤り削減率が得られた．この提案法による認識精度の改善は，従来法に比べ 1%水準で有意な改善である．

謝辞 本研究の一部は文部科学省科学研究費補助金（若手研究 B）課題番号 17700597 による．

参 考 文 献

1) 村上仁一，嵯峨山茂樹：自由発話音声における音響的な特徴の検討，信学論，Vol.J78-D-II, No.12, pp.1741–1749 (1995).
2) Maekawa, K.: Corpus of Spontaneous Japanese: Its Design and Evaluation, *Proc. Workshop Spontaneous Speech Processing and Recognition*, pp.7–12 (2003).
3) 匂坂芳典，東倉陽一：規則による音声合成のた

めの音韻時間制御，信学論，Vol.J67-A，No.7，pp.629–636 (1984).
4) Riley, M.: Tree-based Modelling of Segmental Durations, *Talking Machines: Theories, Models, and Designs*, pp.265–273 (1992).
5) Iwahashi, N. and Sagisaka, Y.: Statistical Modelling of Speech Segment Duration by Constrained Tree Regression, 信学論，Vol.E83-D, No.7, pp.1550–1559 (2000).
6) Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. and Kitamura, T.: Duration Modeling for HMM-Based Speech Synthesis, *Proc. ICSLP*, Vol.2, pp.29–31 (1998).
7) 河井 恒，樋口宣男，山本誠一：基本周波数及び音素持続時間を考慮した音声合成用波形素片データセットの作成，信学論，Vol.J82-D-II, No.8, pp.1229–1238 (1999).
8) 宮武正典，匂坂芳典：種々の発話様式にみられる韻律特徴とその制御，信学論，Vol.J73-D-II, No.12, pp.1929–1935 (1990).
9) 徳田恵一：HMM による音声合成の基礎，信学技報，SP2000, No.74, pp.43–50 (2000).
10) 松尾 広，牧野正三，城戸健一：音素の持続時間モデルに基づく検証法を用いた単語音声認識，信学論，Vol.J73-D-II, No.12, pp.1936–1944 (1990).
11) Molly, L. and Isard, S.: Suprasegmental Duration Modelling with Elastic Constraints in Automatic Speech Recognition, *Proc. ICSLP*, pp.2975–2978 (1998).
12) Anastasakos, A., Schwartz, R. and Shu, H.: Duration Modeling in Large Vocabulary Speech Recognition, *Proc.ICASSP*, pp.628–631 (1995).
13) Levinson, S.: Continusly varible duration hidden Markov models for automatic speech recognition, *Computer Speech and Language*, Vol.1, pp.29–45 (1986).
14) Russell, M.J. and Cook, A.E.: Experimental Evaluation of Duration Modeling Techniques for Automatic Speech Recognition, *Proc. ICASSP*, Vol.4, pp.2376–2379 (1987).
15) Schwartz, R., Austin, S., Kubala, F., Makhoul, J., Nguyen, L., Placeway, P. and Zavaliagkos, G.: New uses for the N-Best sentence hypotheses within the BYBLOS speech recognition system, *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Vol.I, pp.1–4 (1992).
16) 古井貞熙，前川喜久雄，井佐原均：科学技術振興調整費開放的融合研究制度：大規模コーパスに基づく『話し言葉工学』の構築，日本音響学会誌，Vol.56, No.11, pp.752–755 (2000).
17) 匂坂芳典，浦谷則好：ATR 音声・言語データベース，日本音響学会誌，Vol.48, No.12, pp.878–882 (1992).
18) 菊池英明，前川喜久雄：自然発話音声に対する音素自動ラベリング精度の検証，話し言葉の科学と工学ワークショップ，pp.53–58 (2002).
19) Kawahara, T., Lee, A., Kobayashi, T., Takeda, K., Minematsu, N., Sagayama, S., Itou, K., Ito, A., Yamamoto, M., Yamada, A., Utsuro, T. and Shikano, K.: Free Software Toolkit for Japanese large vocabulary continuous speech recognition, *Proc. ICSLP*, Vol.4, pp.476–479 (2000).
20) 伊藤彰則，好田正紀：単語およびクラス n-gram 作成のためのツールキット，信学技報，SP2000, No.106, pp.67–72 (2000).
21) Lee, A., Kawahara, T., Takeda, K. and Shikano, K.: A New Phonetic Tied-Mixture Model For Efficient Decoding, *Proc. ICASSP*, Vol.3, pp.1269–1272 (2000).

(平成 18 年 1 月 16 日受付)
(平成 18 年 9 月 14 日採録)

大河 雄一

平成 10 年東北大学工学部情報工学科卒業．平成 12 年同大学大学院情報科学研究科博士前期課程修了．平成 15 年同大学大学院工学研究科助手．同年同大学大学院教育情報学研究部助手．平成 18 年同大学大学院情報科学研究科博士後期課程修了．博士（情報科学）．音声認識，教育工学等の研究に従事．電子情報通信学会，日本音響学会各会員．

伊藤 彰則（正会員）

昭和 61 年東北大学工学部通信工学科卒業．平成 3 年同大学大学院博士課程修了．同年同大学応用情報学研究センター助手．平成 4 年同大学情報処理教育センター助手．平成 7 年山形大学大学院講師．平成 10〜11 年ボストン大学客員研究員．平成 11 年山形大学工学部助教授．平成 14 年東北大学大学院工学研究科助教授．工学博士．音声言語情報処理，音声信号処理，音楽情報検索等の研究に従事．電子情報通信学会，日本音響学会，IEEE, ISCA 各会員．

鈴木 基之

平成 5 年東北大学工学部情報工学科卒業．平成 7 年同大学大学院工学研究科電気・通信工学専攻博士前期課程修了．同年博士後期課程進学．平成 8 年同課程退学．同年同大学大型計算機センター助手．平成 13 年同大学大学院工学研究科助手．博士（工学）．平成 18〜19 年英国エジンバラ大学客員研究員．音声認識，音声対話，自然言語処理，音楽情報処理の研究に従事．電子情報通信学会，日本音響学会各会員．

牧野 正三（正会員）

昭和 44 年東北大学工学部電子工学科卒業．昭和 49 年同大学大学院博士課程修了．同年同大学電気通信研究所助手．昭和 56 年同大学応用情報学研究センター助手．同助教授同大学大型計算機センター教授，情報シナジーセンター教授を経て，現在同大学大学院工学研究科教授．工学博士．昭和 59〜61 年アメリカ合衆国 STL 客員研究員．言語情報を利用した音声認識・理解の研究，音声データベース，音声信号処理，音声 CALL システム，画像情報処理，文字認識の研究に従事．電子情報通信学会，日本音響学会，人工知能学会，自然言語処理学会，IEEE，ESCA 各会員．