話者照合におけるHMM音声合成による合成音声の判別

佐藤隆之[†,☆] 益子貴史[†]
小林隆夫[†] 徳田恵一[††]

本論文では，HMM音声合成による合成音声を用いた詐称に対して頑健なテキスト指定型話者照合システムについて検討する．ベースラインシステムとして，HMMを用いた連続音素認識に基づくテキスト照合部と，GMMに基づく話者照合部からなるテキスト指定型話者照合システムを用い，これに合成音声判別部を導入する．HMM音声合成システムから生成された音声には，自然音声と比較して話者GMMに対する分析フレームごとの尤度の変動が少ないという特徴がある．そこで，隣接するフレーム間での対数尤度の変動量の絶対値を入力音声全体で平均し，これを合成音声判別のためのパラメータとする合成音声判別部を構築した．実験の結果，各部の閾値をテストデータに対して等誤り率を与える値に設定した場合，合成音声判別部を導入することにより，自然音声の誤り棄却率は0.50%から2.93%へ増加したものの，合成音声の誤り受理率を86.3%から0.69%へ大幅に減少させることができた．

Discrimination of Synthetic Speech Generated by an HMM-based Speech Synthesis System for Speaker Verification

Takayuki Satoh,[†,☆] Takashi Masuko,[†] Takao Kobayashi[†]
and Keiichi Tokuda[††]

This paper describes a text-prompted speaker verification system which is robust to imposture using synthetic speech generated by an HMM-based speech synthesis system. In the verification system, text and speaker are verified separately. Text verification is based on phoneme recognition using HMM, and speaker verification is based on GMM. To discriminate synthetic speech from natural speech, the average inter-frame difference of the log likelihood is calculated, and input speech is judged to be synthetic when this value is smaller than a decision threshold. Experimental results show that the false acceptance rate for synthetic speech was reduced from 86.3% to 0.69%, while the false rejection rate for natural speech was increased from 0.50% to 2.93%.

1. はじめに

音声を用いた本人確認のための手段である話者照合では，申告話者とは異なる話者による成り済まし（詐称）に対する安全性の確保が重要な問題となる．このため，申告話者本人に対する誤り棄却率の低減とともに，詐称者に対する誤り受理率の低減を目指した様々な研究が行われている．しかし，これまでの研究では，人間の声による詐称を対象としたものがほとんどであり，声質変換を用いた詐称についてはいくつか検討されている[1),2)]ものの，合成音声を用いた詐称に関してはほとんど考慮されてこなかった．これは，かつては合成音声の品質があまり高くなかったこと，また任意の話者の声質で音声を合成することが難しかったことなどが理由として考えられる．

一方，近年の音声合成技術の進歩により，高品質な音声や，任意の話者の声質を持つ音声を合成できるようになってきたことも事実である．たとえば，我々の提案する隠れマルコフモデル（HMM）に基づく音声合成システム[3),4)]では，目標となる話者による数文章の発話を用い，合成の基本単位である音素HMMの話者適応を行うことにより，容易に目標話者に近い声質の音声を合成できることを示している[5)]．これらの観点から，我々はこれまでに，HMM音声合成システムを話者照合システムの登録話者に適応することにより，合成音声の話者照合システムによる受理率が非常

† 東京工業大学大学院総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
†† 名古屋工業大学知能情報システム学科
Faculty of Engineering, Nagoya Institute of Technology
☆ 現在，株式会社NTTデータ
Presently with NTT Data Corporation

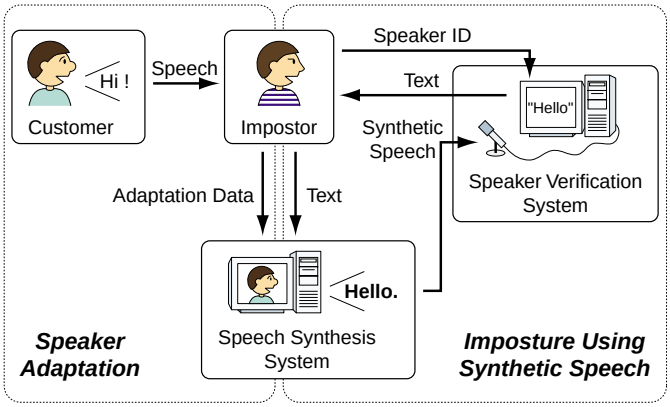


図 1 合成音声を用いた詐称
Fig. 1 An overview of imposture using synthetic speech.

に高くなること，また現在の HMM などの統計的手法に基づく話者照合手法では合成音声と自然音声の判別は難しいことを示した[6]．

そこで本論文では，話者照合システムにおける自然音声と HMM 音声合成システムからの合成音声との判別手法について検討する．ベースラインシステムとして，連続音素認識に基づくテキスト照合部とガウス混合分布（GMM）に基づく話者照合部からなるテキスト指定型話者照合システムを用いる．このベースラインシステムに合成音声判別部を導入し，合成音声に対して安全性の高い話者照合システムを構築することを試みる．

以下，まず 2 章で本論文で仮定している合成音声を用いた詐称の様子を示す．3 章でベースラインとして用いたテキスト指定型話者照合システムについて簡単に述べ，4 章で合成音声と自然音声の判別法について述べる．実験条件および結果を 5 章，6 章に示し，7 章でまとめと今後の課題について述べる．

2. 合成音声を用いた詐称

図 1 に合成音声を用いた詐称の様子を示す．まず，話者照合システムの登録話者の 1 人を成り済ましの目標話者とし，目標話者の少量の音声データが得られたと仮定する．成り済ます相手の音声を得るには，たとえば，相手に直接会って話しかけ音声を録音する，また電話音声を用いた話者照合システムの場合には，成り済ます相手に電話をかけて音声を録音する，などの方法が考えられる．この際，会話の内容はなんでもよく，たとえば道を尋ねる，間違い電話のふりをするなどにより，数十秒程度の音声が得られるような場合を想定している．

HMM 音声合成システムは，合成の基本単位として音素 HMM を用いており，音声認識の分野で研究されている話者適応手法を用いてこの音素 HMM を話者適応することにより，任意の話者に近い声質を持つ音声を合成できるようになる[5]．この HMM 音声合成システムを詐称に用いる場合は，あらかじめ複数の話者のデータベースを用いて学習した音素 HMM を，目標話者から得られた音声を適応データとして話者適応し，目標話者に近い声質を持つ音声を合成できるようにしておく．

照合時には，詐称者は話者照合システムにこの目標話者の話者 ID を入力する．そして，提示されたテキストを音声合成システムに入力し，得られた合成音声を話者照合システムに入力する．話者照合システムは，入力音声が提示したテキストに対応するか，また入力された話者 ID と一致するかを判定し，受理または棄却の判定を行う．合成音声判別部を持つ話者照合システムの場合は，さらに入力音声が自然音声か合成音声かの判定も行う．

3. ベースライン話者照合システム

話者照合手法は，入力音声の発話内容の扱いによって，発話内容によらないテキスト独立型，あらかじめ発話内容が決められているテキスト依存型，照合のつどシステムが発話内容を指定するテキスト指定型に分けることができる[7]．これらのうち，たとえばテキスト独立型の場合には発話内容にかかわらず申告話者本人の音声であれば受理するため，テープレコーダなどによる録音音声を用いた詐称を簡単に許してしまうという欠点がある．そこで本論文では，録音音声による詐称に比較的頑健であると思われるテキスト指定型話者照合システム[8] を対象とし，ベースラインシステムとしてテキスト照合と話者照合をそれぞれ別々に行う

システムを用いる．ベースラインシステムに受理されるためには，テキスト照合部，話者照合部でともに受理される必要がある．

照合時には，まず，指定テキストを話者へ提示し，音声を入力してもらう．入力された音声から特徴パラメータを抽出し，テキスト照合部，話者照合部それぞれへ入力する．

テキスト照合部では，不特定話者音素 HMM を用いた連続音素認識を行い，次式で示される Accuracy を計算する．

$$\text{Accuracy}(\%) = \frac{N - S - D - I}{N} \times 100 \quad (1)$$

ここで，N は提示したテキストに対応する参照音素列に含まれる音素数，S，D，I はそれぞれ置換誤り，脱落誤り，挿入誤りの数を表す．ただし，音素数およびそれぞれの誤り数を求める前にあらかじめ参照音素列および認識により得られた音素列から無音（ポーズ）を取り除いておく．得られた Accuracy が閾値より大きい場合に入力音声が提示したテキストと同一であると判定して受理する．

話者照合部には GMM に基づく話者照合システム[9]を用い，尤度正規化には不特定話者モデルを用いた手法[10]を適用する．学習データから得られた各登録話者の特徴パラメータの分布を話者ごとに GMM でモデル化しておき，またそれとは別に不特定話者 GMM を学習しておく．照合時には，入力音声から得られた特徴パラメータ列 $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T)$ に対して次式で得られる正規化対数尤度 L を求める．

$$L = \frac{1}{T}\left(\sum_{t=1}^{T} \log P(\boldsymbol{o}_t|\lambda_s) - \sum_{t=1}^{T} \log P(\boldsymbol{o}_t|\lambda_{SI})\right) \quad (2)$$

ここで，T は入力音声のフレーム数，λ_s は申告話者 GMM，λ_{SI} は不特定話者 GMM を表し，$P(\boldsymbol{o}|\lambda)$ は特徴パラメータ $\boldsymbol{o}$ に対するモデル λ の尤度である．得られた正規化対数尤度 L を閾値と比較し，L が閾値より大きい場合に入力話者が申告話者と同一であると判定して受理する．

4. 合成音声判別部の導入

話者照合システムの安全性について考えると，他の話者による音声を棄却できるだけではなく，合成音声を棄却できる必要がある．しかし，これまでに様々な音声合成手法が提案されており，合成音声の持つ特徴や品質も様々であることから，すべての合成音声について考慮することは容易ではない．そこで本論文では，すべての合成音声を対象とはせず，これまでに話者照合システムによる受理率が高いことが示されている[6]HMM 音声合成システム[3),4)]からの合成音声について検討する．

HMM 音声合成システムでは，音声の特徴パラメータの統計量を HMM を用いて学習し，合成時には尤度最大基準に基づいて特徴パラメータ列を生成している．このため，従来の GMM や HMM などの統計モデルの尤度と閾値との比較に基づく話者照合手法[8),9)]では，自然音声との判別は難しいと考えられる．実際，文献 6) では，わずか 1 文章の音声を用いて登録話者に HMM 音声合成システムを適応した場合でも合成音声の誤り受理率と自然音声の誤り棄却率との等誤り率が 30% 以上となること，また話者によっては合成音声の方が自然音声よりも正規化対数尤度が高くなる場合があることを示している．このことから，従来の GMM や HMM を用いた話者照合手法では，効率的に合成音声を棄却することはできないことが分かる．そこで，自然音声と HMM 音声合成システムによる合成音声との違いに着目した新たな合成音声判別部を構築し，話者照合システムに導入する．

図 2 (a) に自然音声を分析して得られたスペクトル，(b) に HMM 音声合成システムから生成されたスペクトルの例を示す．発話内容は「六百人…」で，自然音声のスペクトル分析は 5.2 節に示す音声合成システムにおける分析条件に従って行った．図 2 (a) と (b) を比べると，合成音声のスペクトルの方が自然音声から得られたものよりも滑らかに変化していることが分かる．これは，スペクトルパラメータ生成時に静的特徴量と動的特徴量をともに考慮し，HMM から得られる出力分布列に対して尤度が最大となるパラメータ列を求めているため，自然音声にみられるようなランダムな変動が起こらないためである．そこで，この特徴を自然音声と合成音声との判別に利用する．

まず，入力音声から得られた特徴パラメータ列 $\boldsymbol{O} = (\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T)$ に対し，隣接するフレーム間での対数尤度の変動量の絶対値 Δl_t を次式のように定義する．

$$\Delta l_t = |l_t - l_{t-1}| \quad (3)$$

ここで l_t はフレーム t の特徴パラメータ $\boldsymbol{o}_t$ に対する申告話者モデル λ_s の対数尤度

$$l_t = \log P(\boldsymbol{o}_t|\lambda_s) \quad (4)$$

である．HMM 音声合成システムから生成された特徴パラメータの時間変動は自然音声よりも小さいため，合成音声から求められた Δl_t の値は自然音声よりも小さくなる傾向があると考えられる．そこで，全音声

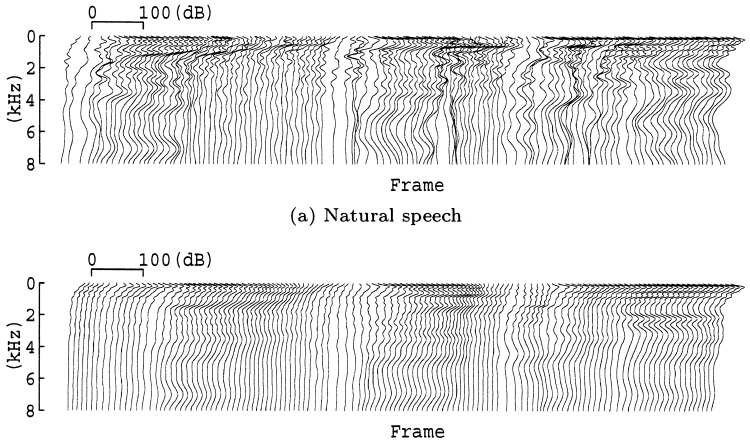


図 2 自然音声と合成音声のスペクトル（発話内容「六百人…」）
Fig. 2 Spectra obtained from natural speech and generated by HMM-based speech synthesis system (/r-o-cl-py-a-k-u-n-i-N/).

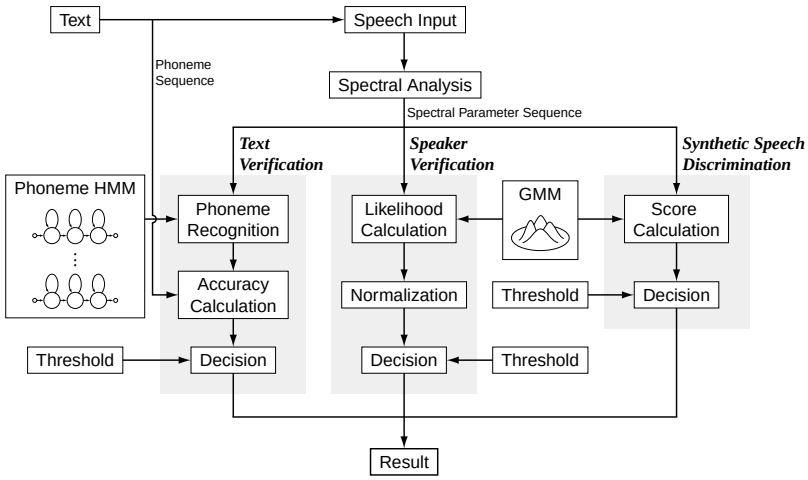


図 3 合成音声判別部を導入した話者照合システムのブロック図
Fig. 3 Blockdiagram of a robust speaker verification system against HMM-based speech synthesis.

区間の Δl_t の平均値

$$D = \frac{1}{T-1}\sum_{t=2}^{T}\Delta l_t \qquad (5)$$

を合成音声判別のためのパラメータとして用いる．合成音声判別部では，平均値 D が閾値より小さい場合，入力音声が合成音声であると判定して棄却する．

図 3 にベースラインシステムに合成音声判別部を加えた提案システムのブロック図を示す．提案システムは，テキスト照合部（Text Verification），話者照合部（Speaker Verification）と，合成音声判別部（Synthetic Speech Discrimination）の 3 つの部分からなる．提案システムに受理されるためには，テキスト照合部，話者照合部でそれぞれ受理されることに加え，合成音声判別部で自然音声であると判定されることが必要である．

5. 実験条件

表 1 に話者照合システムおよび音声合成システムで利用した GMM および HMM の学習，適応に用いた話者数と文章数を示す．音声合成システムの学習に用いた話者は，話者照合システムの登録話者と重複がないように選んでいる．また，音声合成システムの学習，適応に用いた文章についても，登録話者 GMM の

表 1 学習データ
Table 1 Training data.

		話者	文章
登録話者 GMM		1 名（登録話者毎）	5 文章
不特定話者 GMM		30 名	4,500 文章
テキスト照合用 HMM		132 名	20,414 文章
音声合成用 HMM	学習	6 名	2,118 文章
	適応	1 名（登録話者毎）	5 文章

表 2 申告話者，提示テキストの組に対する入力音声
Table 2 Test data for a combination of claimed speakers and prompted sentences.

	話者	文章
登録話者	16 名	100 文章
登録外話者	20 名	100 文章
合成音声	1 名（申告話者のみ）	100 文章

学習およびテストに用いた文章と重複がないように選んでいる．

5.1 話者照合システム

ATR 音声データベース[11] セット C（多数話者データベース）から話者照合システムの登録話者として男性話者 16 名，システムに入力する登録外の話者として男性話者 20 名を用い，各登録話者ごとに 5 文章の発話を用いて GMM を学習した．また，尤度正規化に用いる不特定話者 GMM の学習には，日本音響学会研究用連続音声データベースの男性話者 30 名による 4,500 文章の発話を用いた．GMM の混合数はそれぞれ 32 とした．

テキスト照合には，IPA 日本語ディクテーション基本ソフトウェア[12),13)] に収録されている男性用 3 状態 16 混合 monophone HMM を用いた．これらの HMM は，男性話者 132 名によるのべ 20,414 文章の発話を用いて学習されたものである．音素数は無音を含めて 43 である．

話者照合システムにおける入力音声のサンプリング周波数は 16 kHz，分析窓はフレーム長 25 ms のハミング窓，フレーム周期は 10 ms とし，特徴パラメータは 1〜12 次メル周波数ケプストラム係数（MFCC），ΔMFCC および Δ パワーとした．また，周波数分析におけるフィルタバンク数は 24 とし，発話ごとにケプストラム平均正規化を行った．

5.2 音声合成システム

音声合成システムは，ATR 音声データベースセット B（連続音声データベース）の男性話者 6 名による合計 2,118 文章の発話を用いて学習し，話者照合システムの各登録話者に対して 5 文章の発話を用いて適応[5] した．

音声合成システムにおける音声のサンプリング周波数は 16 kHz，分析周期は 5 ms とした．スペクトル分析にはフレーム長 25 ms のブラックマン窓を用いた 24 次メルケプストラム分析[14]，ピッチ抽出には ESPS[15] に含まれる *get_f0* プログラムを用い，特徴パラメータは 0〜24 次メルケプストラム，対数基本周波数，およびそれらの Δ，Δ^2 パラメータとした．HMM は 5 状態 left-to-right モデルで，スペクトル部は単一対角共分散ガウス分布，ピッチ部，Δ ピッチ部，Δ^2 ピッチ部はそれぞれ多空間確率分布[16] でモデル化した．音素数は無音を含めて 42，各音素モデルは音韻，韻律に影響を与える変動要因を考慮したコンテキスト依存モデルとし，最小記述長（MDL）基準を用いた決定木に基づくコンテキストクラスタリング[17] により状態の共有を行っている．また，HMM の各状態はガウス分布で近似された状態継続長分布を持っており，スペクトル，ピッチの分布と同様にクラスタリングされている[4]．合成時には，まず，状態継続長分布に従って各状態の継続長を決定する．そして，尤度最大化基準に基づくパラメータ生成アルゴリズム[18] によりメルケプストラム列およびピッチパターンを生成し，メル対数スペクトル近似（MLSA）フィルタ[19] を用いて音声を合成する．

5.3 評価方法

表 2 に，照合時の申告話者および指定テキストの各組合せに対して入力した音声データの話者数および文章数を示す．

照合時に指定するテキストとして，ATR 音韻バランス文より，登録話者 GMM の学習，音声合成システムの学習および適応に用いた文章と重なりのない 100 文章を用いた．照合時には，各登録話者，指定テキストの組合せに対し，自然音声の場合はすべての話者（申告話者 1 名，申告話者以外の登録話者 15 名，登録外話者 20 名の合計 36 名）によるすべてのテスト文章の発話を照合システムへの入力とし，合成音声の場合は申告話者に適応したモデルによるすべてのテスト文章に対応する合成音声を照合システムへの入力とした．

提案する話者照合システムでは，テキスト照合部，話者照合部，および合成音声判別部のすべてで受理された場合のみ入力音声を申告話者の音声であると判定して受理するため，システム全体の性能は各部の閾値の定め方により大きく影響を受ける．しかし，閾値を適切に設定することは難しく，また安全性と利便性のどちらを優先するかによっても閾値の定め方が異なってくる．そこで，システム全体の性能を示す 1 つの目安として，テキスト照合部，話者照合部の閾値は自然音声に対して等誤り率（EER）を与える値に，また合

成音声判別部は合成音声と自然音声に対して EER を与える値にそれぞれ設定した場合の誤り棄却率および誤り受理率を求めた．

6. 結果と考察

6.1 ベースラインシステムの評価

図 4 に話者照合部における自然音声に対する誤り棄却率（FRR）と誤り受理率（FAR），および合成音声に対する FAR を示す．図の横軸は正規化対数尤度に対する閾値，縦軸は誤り率を表し，実線は申告話者による自然音声に対する FRR，点線は申告話者以外の話者による自然音声に対する FAR，一点鎖線は合成音声に対する FAR をそれぞれ表す．

自然音声に対する EER は 0.46% となったが，閾値を自然音声に対する EER を与える値（図中の破線で示す）に設定した場合，合成音声に対する FAR は 86% 以上となった．また，自然音声と合成音声の EER は 27% 以上となっていることから，閾値の調整のみでは合成音声と自然音声を効率的に判別することは難しいことが分かる．

図 5 にテキスト照合部における自然音声に対する FRR と FAR を示す．図の横軸は Accuracy に対する閾値，縦軸は誤り率を表し，実線は指定テキストの読み上げ音声に対する FRR，点線は指定テキストと異なる音声に対する FAR をそれぞれ表す．

自然音声に対するテキスト照合では EER は 0.028% となったことから，テキスト照合部は十分な性能を持っていると考えられる．また，閾値を自然音声に対して EER を与える値に設定した場合，合成音声に対する FRR と FAR はそれぞれ 0.062%，0.043% となった．ただし，合成音声の場合でも自然音声と同様，指定テキストに対応する合成音声を入力したときに棄却された場合，および指定テキストと異なる合成音声を入力したときに受理された場合をそれぞれ誤りとしている．

話者照合部，テキスト照合部の閾値を，テストデータに対する等誤り率を与える値（図 4，図 5 の破線で表される値）に設定した場合の，ベースラインシステム全体での自然音声に対する FRR と FAR，および合成音声に対する FAR を**表 3** に示す．自然音声に対しては，FRR，FAR ともに低い値となっているが，合成音声に対する FAR は 86% 以上となっており，HMM 音声合成システムを用いた詐称が十分可能であることが分かる．

6.2 合成音声判別部およびシステム全体の評価

図 6 に自然音声と合成音声から得られた Δl_t の例を示す．横軸はフレーム，縦軸は Δl_t の値を表し，破線は自然音声から，実線は合成音声からそれぞれ求められた Δl_t を表す．図より，実際に合成音声から求められた Δl_t の値は自然音声よりも小さくなる傾向にあることが分かる．

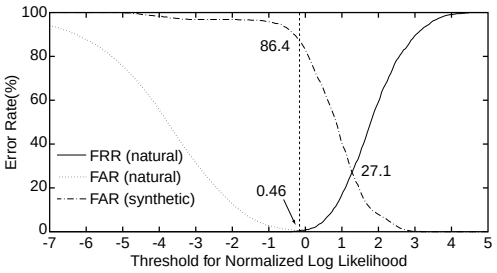


図 4 話者照合部の誤り率

Fig. 4 False acceptance and rejection rates of speaker verification part as a function of the values of the decision threshold.

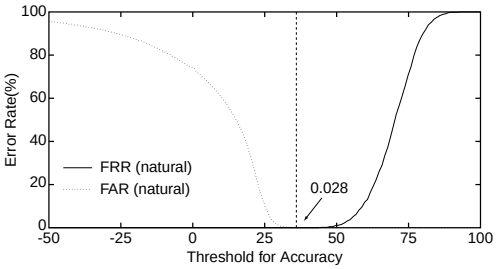


図 5 テキスト照合部の誤り率

Fig. 5 False acceptance and rejection rates of text verification part as a function of the values of the decision threshold.

表 3 ベースラインシステム全体での誤り率

Table 3 The false rejection and acceptance rates (%) for the speaker verification system without the synthetic speech discrimination part.

natural		synthetic
FRR	FAR	FAR
0.50	0.005	86.3

図 7 に合成音声判別部の FRR および FAR を示す．図の横軸はフレーム間対数尤度変動量の平均値に対する閾値，縦軸は誤り率を表し，実線は自然音声に対する FRR，点線は合成音声に対する FAR をそれぞれ表す．EER は 2.5% となっており，提案手法による自然音声と合成音声との判別が十分可能であることが分かる．

合成音声判別部を導入した話者照合システム全体での自然音声に対する FRR と FAR および合成音声に対する FAR を**表 4** に示す．合成音声の判別部の閾値は，テストデータに対する等誤り率を与える値（図 7

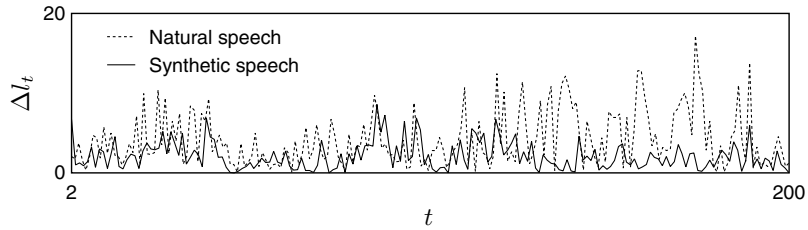


図 6 隣接フレーム間での対数尤度の変動量
Fig. 6 Inter-frame difference of log likelihood for natural and synthetic speech.

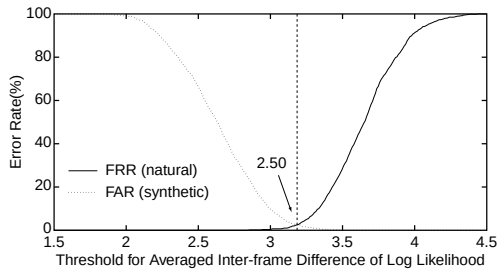


図 7 合成音声判別部の誤り率
Fig. 7 False acceptance and rejection rates of synthetic speech discrimination part as a function of the values of the decision threshold.

表 4 合成音声判別部を持つ話者照合システム全体の誤り率
Table 4 The false rejection and acceptance rates (%) for the speaker verification system with the synthetic speech discrimination part.

natural		synthetic
FRR	FAR	FAR
2.93	0.004	0.69

の破線で表される値）に設定した．表 3 と比較して，自然音声の FRR は多少上昇したものの，合成音声の FAR を大幅に低下させることができたことが分かる．

6.3 他の音声合成システムに対する効果

本論文で提案する合成音声の判別手法は，合成音声のスペクトルの時間変化の滑らかさを利用した手法であることから，ボコーダ型でスペクトルを滑らかに補間している音声合成システムに対しては，有効であると考えられる．逆に，波形素片接続型の音声合成システムの場合には，素片内では自然音声と同様にスペクトルがランダムに変動し，素片接続部では自然音声よりも大きくスペクトルが変化することから，提案手法で判別することは難しいと考えられる．しかし，波形素片接続型の音声合成システムで合成音声の声質を変えるためには素片データベースを再構築する必要があり，本論文で想定しているように目標話者の音声がわずかしか得られない場合には，十分な素片データベースを構築することができないため詐称に用いることは難しいと考えられる．また，その他の音声合成システムについても詐称に用いることが可能であるか必ずしも明らかではない．他の音声合成システムに対する有効性の検討は今後の課題となる．

7. ま と め

本論文では，話者照合システムのための HMM 音声合成システムからの合成音声と自然音声の判別法を提案した．実験結果より，合成音声の誤り受理率を大幅に低下できることが示され，提案手法の有効性が示された．提案手法は合成音声のスペクトルの滑らかさを利用した手法であるため，HMM 音声合成システムによる合成音声のみではなく，他の滑らかなスペクトルを持つ合成音声に対しても有効であると考えられる．そこで，他の合成音声に対する提案手法の有効性の検討が今後の課題としてあげられる．また，提案手法が有効ではないと考えられる波形素片接続による合成音声の判別法の検討も今後の課題となる．

謝辞 研究を進めるにあたり，実験にご協力をいただいた東京工業大学大学院博士課程田村正統氏に感謝します．本研究の一部は文部省科学研究費補助金（奨励研究（A）課題番号 11750311）によった．

参 考 文 献

1) Genoud, D. and Chollet, G.: Speech preprocessing against intentional imposture in speaker recognition, *Proc. ICSLP-98*, pp.105–108 (1998).
2) Pellom, B.L. and Hansen, J.H.L.: An experimental study of speaker verification sensitivity to computer voice-altered imposters, *Proc. ICASSP-99*, pp.837–840 (1999).
3) 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（D-II），Vol.J79-D-II, No.12, pp.2184–2190 (1996).
4) 吉村貴克，徳田恵一，益子貴史，小林隆夫，北村正：HMM に基づく音声合成におけるスペクトル・ピッチ・継続長の同時モデル化，信学論（D-II），

Vol.J83-D-II, No.11, pp.2099–2107 (2000).
5) Tamura, M., Masuko, T., Tokuda, K. and Kobayashi, T.: Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR, *Proc. ICASSP-2001*, pp.805–808 (2001).
6) 益子貴史，徳田恵一，小林隆夫：話者照合システムに対する合成音声による詐称，信学論（D-II），Vol.J83-D-II, No.11, pp.2283–2290 (2000).
7) 古井貞熙：音声情報処理，森北出版 (1992).
8) 松井知子，古井貞熙：テキスト指定型話者認識，信学論（D-II），Vol.J79-D-II, No.5, pp.647–656 (1996).
9) Reynolds, D.A.: Speaker identification and verification using Gaussian mixture speaker models, *Speech Communication*, Vol.17, No.1–2, pp.91–108 (1995).
10) Matsui, T. and Furui, S.: Likelihood normalization for speaker verification using a phoneme- and speaker-independent model, *Speech Communication*, Vol.17, No.1–2, pp.109–116 (1995).
11) 株式会社国際電気通信基礎技術研究所開発センタ．http://www.red.atr.co.jp/
12) http://winnie.kuis.kyoto-u.ac.jp/dictation/
13) 鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹男：音声認識システム，オーム社出版局 (2001).
14) 徳田恵一，小林隆夫，深田俊明，斎藤博徳，今井聖：メルケプストラムをパラメータとする音声のスペクトル推定，信学論（A），Vol.J74-A, No.8, pp.1240–1248 (1991).
15) Entropic Research Laboratory, Inc.: *ESPS Programs Version 5.0* (1993).
16) 徳田恵一，益子貴史，宮崎　昇，小林隆夫：多空間上の確率分布に基づいた HMM, 信学論（D-II），Vol.J83-D-II, No.7, pp.1579–1589 (2000).
17) Shinoda, K. and Watanabe, T.: MDL-based context-dependent subword modeling for speech recognition, *The Journal of the Acoustical Society of Japan* (*E*), Vol.21, No.2, pp.79–86 (2000).
18) 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，日本音響学会誌，Vol.53, No.3, pp.192–200 (1997).
19) 今井　聖，住田一男，古市千枝子：音声合成のためのメル対数スペクトル近似（MLSA）フィルタ，信学論（A），Vol.J66-A, No.2, pp.122–129 (1983).

(平成 13 年 11 月 15 日受付)
(平成 14 年 4 月 16 日採録)

佐藤　隆之

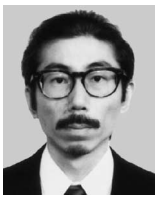
昭和 52 年生．平成 13 年東京工業大学大学院総合理工学研究科物理情報システム創造専攻修士課程修了．在学中，話者認識の研究に従事．日本音響学会会員．

益子　貴史

昭和 45 年生．平成 7 年東京工業大学大学院総合理工学研究科知能科学専攻修士課程修了．同年東京工業大学精密工学研究所助手．平成 11 年同大学大学院総合理工学研究科物理情報システム創造専攻助手．音声情報処理，マルチモーダルインタフェースの研究に従事．平成 12 年度電子情報通信学会論文賞（第 57 回），同猪瀬賞（第 7 回）受賞．日本音響学会，電子情報通信学会，IEEE, ISCA 各会員．

小林　隆夫（正会員）

昭和 30 年生．昭和 57 年東京工業大学大学院博士課程修了．同年東京工業大学精密工学研究所助手．同助教授を経て現在東京工業大学大学院総合理工学研究科物理情報システム創造専攻教授．工学博士．ディジタルフィルタ，音声分析・合成・符号化・認識，マルチモーダルインタフェースの研究に従事．平成 12 年度電子情報通信学会論文賞（第 57 回），同猪瀬賞（第 7 回），電気通信普及財団賞受賞．日本音響学会，電子情報通信学会，IEEE, ISCA 各会員．



徳田　恵一（正会員）

昭和 35 年生．平成元年東京工業大学大学院博士課程修了．同年東京工業大学電気電子工学科助手．平成 8 年名古屋工業大学知能情報システム学科助教授．工学博士．音声分析・合成・符号化・認識，ディジタル信号処理，マルチモーダルインタフェースの研究に従事．平成 12 年度電子情報通信学会論文賞（第 57 回），同猪瀬賞（第 7 回）受賞．日本音響学会，電子情報通信学会，人工知能学会，IEEE 各会員．