G-009

ピッチの時間変化パターンを用いた合成音声判別法

Discrimination Method of Synthetic Speech Using Pitch Time-Varying Pattern

荻原 昭夫† 海野 仁‡ 汐崎 陽†
Akio Ogihara Hitoshi Unno Akira Shiozaki

1．まえがき

音声に関する研究の成果が実社会に応用されつつある．話者照合の分野においても，「機密性の高い研究室への入室管理」や「コンピュータシステムのアクセス管理」などに音声による本人認証が導入されている．現状の話者照合システムは，悪意ある第三者の声帯模写による詐称行為（正規利用者になりすます行為）に対しても対処可能である．しかしながら，近年，合成音声を用いた詐称行為により話者照合システムを不正に突破可能である事が報告[1]されている．したがって，今後は合成音声を用いた詐称への対応が不可欠になると考えられる．

そこで我々は，図 1 に示すように，従来の話者照合システムの前段に「合成音声判別システム」を配置することで，事前に合成音声を排除することを考えている．なお，合成音声判別システムは話者照合システムとは独立しているので，既存の話者照合システム自体に修正を施す必要は無く，合成音声判別システムを追加するだけで合成音声への対応が可能である．

本論文では，人間により発声された自然発声と合成音声ではピッチ（基本周波数）の変化パターンに差異がある事に着目し，ピッチの時間変化パターンを用いて合成音声の判別を行なう手法を提案する．

2．ピッチの時間変化パターン

合成音声の判別を行うために，短区間自己相関関数[2]を使用して，音声信号から「ピッチの時間変化パターン」を以下の手順で求める．

まず，音声信号 $x(t)$に対する分析時刻 t における遅れ時間τの短区間自己相関関数 $R(t,\tau)$を次式により求める．

$$R(t,\tau)=\int_{-l(\tau)/2}^{l(\tau)/2} x(t+\xi-\tau/2)x(t+\xi+\tau/2)d\xi \qquad (1)$$

なお，$l(\tau)=(m-1)\tau$，m は 2 以上の整数と定義する．式(1)の積分期間はτに応じて変化し，積分区間が長いほど $R(t,\tau)$の値は大きくなる．そこで，式(2)に従って正規化を施すことで積分期間の影響を排除する．なお，式中の分母 $P(t,\tau)$は式(3)により求められ，これは式(1)を算出する際に使用した音声信号のパワーに相当する．

$$\phi(t,\tau)=\frac{R(t,\tau)}{P(t,\tau)} \qquad (2)$$

$$P(t,\tau)=\frac{1}{2}\left\{\int_{-l(\tau)/2}^{l(\tau)/2} x(t+\xi-\tau/2)^2 d\xi+\int_{-l(\tau)/2}^{l(\tau)/2} x(t+\xi+\tau/2)^2 d\xi\right\} \qquad (3)$$

† 大阪府立大学大学院工学研究科, Graduate School of Engineering, Osaka Prefecture University

‡ 警察庁, National Police Agency

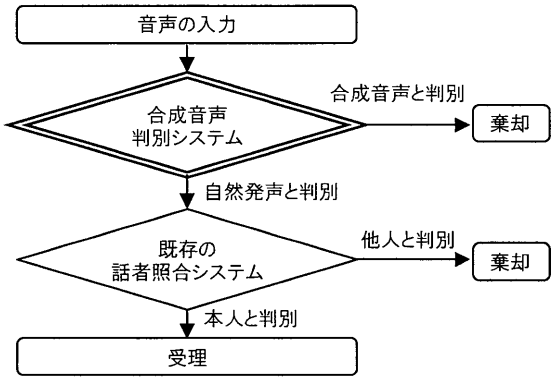


図 1　合成音声判別システムの追加による合成音声詐称への対応

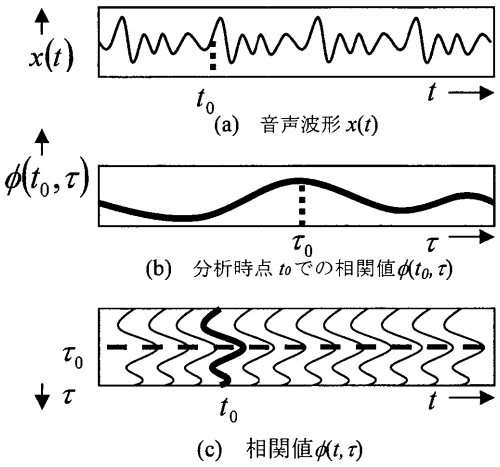


図 2　ピッチの時間変化パターンの作成過程

式(2)で求められる相関値$\phi(t,\tau)$から「ピッチの時間変化パターン」を得る過程を図 2 に示す．図 2(a)の音声波形に対して，ある分析時刻 t_0における$\phi(t_0,\tau)$を求めたのが図 2 (b)である．次に，相関値の算出をすべての分析時刻 t で行なうことで，図 2 (c)に示すような三次元グラフが得られる（図 2 (c)の太線は図 2 (b)のグラフに対応している）．この三次元グラフを「ピッチの時間変化パターン」とみなす．合成音声と自然発声のピッチの時間変化パターンの一例を図 3 に示す．なお，図中では相関値の大きさに比例して画素の輝度を明るく表示することにより，三次元グラフを擬似的に表現している．

3．合成音声判別法

図 3 に示したピッチの時間変化パターンを観察すると，“相関値$\phi(t, \tau)$の極大点の位置の時間的変化”や“分析時刻 t にともなう高輝度部分（白い部分）の面積変化”に関して，合成音声と自然発声で差異が生じている．これらの特徴を数値化することで，合成音声と自然発声の判別が可能と考えられる．

第一の特徴量として，相関値の極大点を表す「ピーク点の位置」が考えられる．次に，高輝度部分と低輝度部分との境界線を意味するものとして，「上半値点の位置」と「下半値点の位置」を第二および第三の特徴量として考える．なお“半値点の位置”は，同一分析時刻における相関値のピーク位置を始点として上下方向に走査し，相関値の値が“ピーク値の 50%の値”となる点の位置とした．第四の特徴量として，高輝度部分の面積を表す「半値幅」を考える．半値幅は“上半値点の位置”と“下半値点の位置”との距離である．例として，図 3 から抽出したピーク点の位置，上半値点の位置，下半値点の位置，半値幅を図 4 に示す．

合成音声の判別に先立って，話者照合システムの正規利用者の音声信号から上述の特徴量を抽出し，各特徴量の時系列を登録データとして合成音声判別システムに事前に登録しておく．そして，実際に判別を行なう際には，“入力音声から抽出した特徴量の時系列”と“登録データ”との距離を DP マッチングにより求め，距離が閾値より大きい場合は「現在入力されている音声は合成音声である」と判別する．

4．実験結果

提案手法の有効性を確認するために，以下の条件で合成音声の判別実験を行なった．

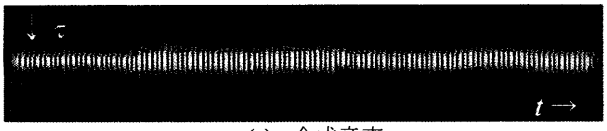
自然発声のサンプルとして，母音により構成される単語「あいう」を成人男性が 100 回発声した音声を使用した．さらに，これらの自然発声サンプルをもとに，文献[1]の手法によって話者照合システムに対して高い詐称能力を有する合成音声を作成し，これらを合成音声のサンプルとして使用した．なお，サンプル数は自然発声，合成音声ともに 100 個である．

式(1)および式(3)において m=2 として計算を行ない，ピッチの時間変化パターンを求めた．四種類の特徴量（ピーク点の位置，上半値点の位置，下半値点の位置，半値幅）のそれぞれについて，DP マッチングにより“自然発声同士の距離”および“合成音声と自然発声の距離”を求めた．一例として，特徴量として「半値幅」を用いた場合の実験結果を図 5 に示す．二つの分布の交点における距離を閾値とした場合，合成音声の約 99%を正しく判別することができた．なお，同様の基準で閾値を設定した場合，「ピーク点の位置」では 93%，「上半値点の位置」では 96%，「下半値点の位置」では 98%の判別成功率が得られ，今回用いた特徴量の中では「半値幅」が最も合成音声の判別に有効である事が分かった．

5．むすび

本論文では，ピッチの時間変化パターンを用いて合成音声の判別を行なう手法を提案した．四種類の特徴量について比較実験を行なった結果，「半値幅」が最も有効であり，99%の精度で合成音声を判別可能であった．

今後は，“新たな特徴量の導入”や“複数の特徴量の組み合わせ”等を検討することで，合成音声の判別精度を高める予定である．



(a) 合成音声

(b) 自然発声

図 3 ピッチの時間変化パターン

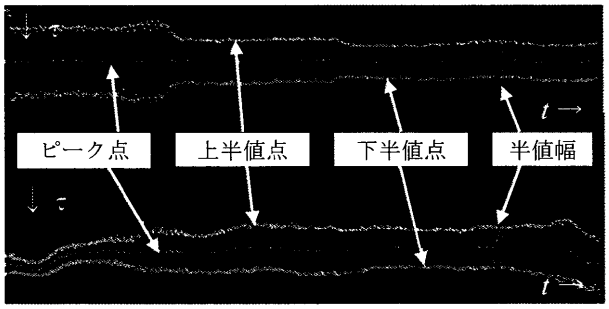


図 4 ピッチの時間変化パターン上の特徴量
（上：合成音声，下：自然発声）

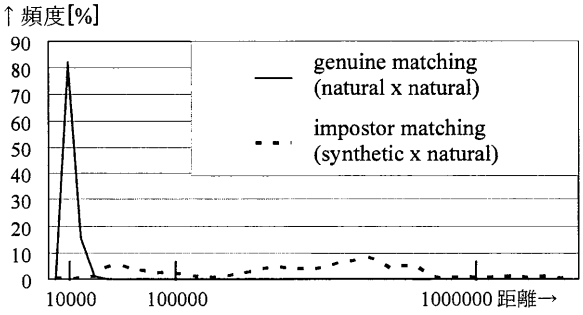


図 5 特徴量として半値幅を用いた場合の実験結果
（実線：自然発声同士の距離の分布，
点線：合成音声と自然発声との距離の分布）

参考文献

[1] 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, Nov. 2000.

[2] 藤崎博也，広瀬啓吉，瀬戸重宣，“分析窓位置による誤りの少ない音声ピッチ抽出方式，”信学技報, SP89-69, pp.1-8, Nov. 1989.