内容梗概

近年, 外国語教育の支援を目的とした CALL(Computer–Assisted Language Learning) システムの需要が高まり, 広く用いられるようになってきた．これらの CALL システムの利用する音声認識技術のほとんどが，スペクトルを用いて音声を表象することに基づいている．この場合音声は, 発音の良し悪しだけでなく，年齢や性別といった，発音とは無関係の要因によっても音声表象は変形してしまう．このため「不一致問題」と呼ばれる, 利用者の身体的特徴により発音評価の結果が不安定になる問題が起こる．現在の音声認識システムでは，年齢や性別などの話者の違いに起因する音声表象の変形に対する頑健性を確保すべく，多数話者の音声を収集し，発声を統計的にモデル化する技術（統計的音響モデル）や，必要に応じて音響モデルを利用者の声色に適応させる技術を用いて, この不一致問題の解決を図っているが，スペクトルを用いて音声を表象していることに変わりはなく，根本的な解決には至っていない．

この不一致問題に対する抜本的解決は，音声に含まれる非言語的な情報のみを除去することである．近年, この非言語的情報をそぎ落した上で発音を表現する「音声の構造的表象」が提案され, この表象を用いて教師と学習者の音声を「発音の良し悪し」のみで比較する CALL システムの構築が行われている．音声の構造的表象に基づくため，これらの CALL システムでは比較する音響イベントの対応は既知である必要がある．既に構築された CALL システムでは，発音の対応から音素同士の対応を明確にしたり，多量の発声を用いて話者の音響モデルを連結学習によって明確にしている．

既存のシステムでは，特定の音素や単語に依存した比較が行われていたが，著者が目標とするのは，構造的表象を用いてフレーズ発声などのより一般的な発音を比較するシステムであり，そのために二発声間の時間アライメントを精度よく求めることが必要である．本研究では，話者性の異なる教師・学習者それぞれの一発声から二発声間における時間アライメント手法に関して検討をおこなった．