実環境音声処理――音声認識に適した残響除去収音

音声は，多くの人にとって，もっとも自然で使いやすいコミュニケーション手段の１つです．私たちは，高速高精度な音声認識技術を中核に，快適で安心な音声インタフェース技術の実現・高度化を目指した研究開発を続けています．本稿では，音声認識の前処理である収音技術のうち，特に，残響除去収音技術について，最近の研究開発状況を紹介します．

木下（きのした） 慶介（けいすけ） /中谷（なかたに） 智広（ともひろ）
三好（みよし） 正人（まさと）
NTTコミュニケーション科学基礎研究所

快適で安心な音声インタフェース

私たちは，音声インタフェースを，来るべきブロードバンド・ユビキタス社会において，快適かつ安心に，インフォメーション・ネットワークへアクセスするための最重要技術の１つに位置付けています．

音声インタフェース技術は，中核となる高速高精度な音声認識技術と，音声強調収音技術とから構成されます．音声認識は，音声発話を文字列に変換する技術です．私たちの音声認識システムは，日本語講演音声を収録した日本語話し言葉コーパス（CSJ: Corpus of Spontaneous Japanese）を用いた性能評価において，現時点における最高クラスの性能を達成しています[(1)]．音声強調収音は，騒音や他の人の話し声，残響など，目的音声の品質を低下させ，音声認識の性能を低下させる不要音をマイクロホン収音された音声から取り除く技術[(2)]です．これまでに，たくさんの話し声や騒音の中から，マイクロホンに比較的近い音声を分離し，強調することにも成功しています．

本稿では，残響除去収音技術について，最近の研究開発状況を紹介します．

音声の残響歪みと音声認識性能

残響とは，話者から離れたところにあるマイクロホンで音声収音する際，直接音に続いてマイクロホンに到達する多数の反射音を指します．**図１**は反射音がマイクロホンに時々刻々と到来する様を表した模式図です．これらの反射音の中でも，とりわけ，直接音から少し（30～50ミリ秒以上）遅れてマイクロホンに到達する後部残響は，音声認識性能を大きく低下させる要因となります．

音声認識は，マイクロホン収音された音声を，あらかじめ学習した音声のモデルとパターン照合し，文字列に変換する技術です．収音音声が残響で歪むと，パターン照合に失敗することが多くなり，音声認識性能は低下します．

この問題への対策として，音声認識に用いる音声モデルを残響環境下で学習する方法や，音声認識の前処理として，収音音声から残響歪みを取り除く Cepstral Mean Subtraction（CMS）などの方法が検討されてきました．例えば，CMSを用いると，音声認識で用いられる30ミリ秒程度の音声の分析窓長内に収まる，初期反射音（図１）の影響は十分に抑圧できます．しかし，他の従来法と同様に，音声の分析窓長を超える後部残響による歪みを低減することは困難でした．

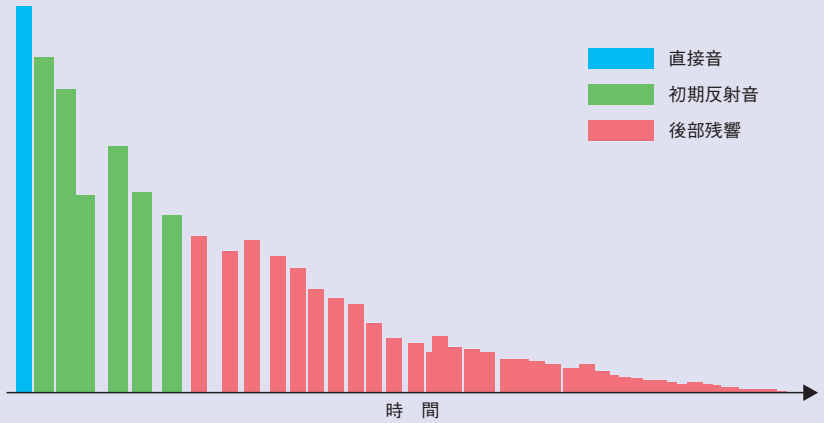

図１　反射音到来の模式図

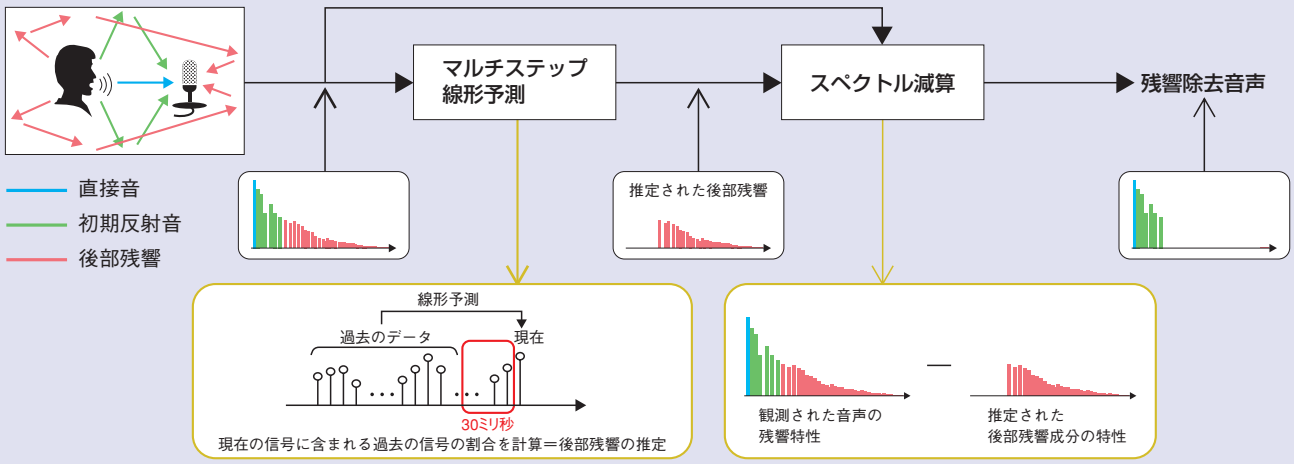

図２　提案方法の処理フロー

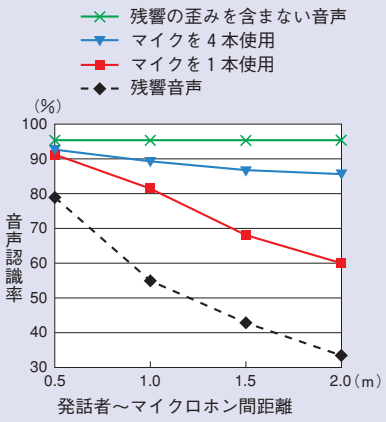

図３　音声認識による評価結果

私たちは，マイクロホン収音された音声に含まれる後部残響歪みを取り除くため，マルチステップ線形予測に基づく新しい残響除去方法を提案しました．

新しい残響除去方法

提案法の処理フローを**図２**に示します．まず，マルチステップ線形予測を用いて，マイクロホン収音された音声に含まれる後部残響歪みを計算します．マルチステップ線形予測は，現在時刻の収音音声に含まれる，30ミリ秒以上前の時刻の信号の割合を計算する統計的な手法です．この割合が後部残響歪みに相当します．次に，スペクトル減算を用いて，計算した後部残響歪みを収音音声から差し引きます．最後に，CMSを用いて，初期反射音の影響も取り除きます．以上の結果，収音音声に含まれる残響歪みは低減されます．

提案方法の残響除去効果を調べるために行った音声認識実験の結果を**図３**に示します．0.5秒の残響時間を持つ4.5 m×3.5 m×2.5 mの実験室を用いました．音響条件は，10畳のフローリング床の洋間と同じくらいです．

図３において，横軸は発話者～マイクロホン間距離を，縦軸は音声認識性能（認識率）をそれぞれ表します．黒点線は残響歪みを含む音声（残響音声）の認識率を表します．赤線と青線は，ともに提案方法の残響除去効果を表し，赤線はマイクロホンを１本，青線はマイクロホンを４本用いた場合をそれぞれ表します．緑線は残響歪みを含まない音声の認識性能を表します．

残響音声の認識性能は，発話者～マイクロホン間距離の増加に従い大幅に低下します．一方，提案法で処理された残響音声の認識性能は，それほど低下しません．発話者～マイクロホン間距離が0.5 m程度の場合は，１本のマイクロホンで，十分な音声認識性能を得ました．また，マイクロホンを４本に増やすと，音声認識性能はさらに改善されました．

今後の展開

上記提案方法は，後部残響歪みの計算に数秒のマイクロホン収音音声を用いています．今後は，計算に必要な音声データ量の削減と，逐次型処理の導入を図り，残響歪みを実時間で低減できる方法へと発展させたいと思います．

■参考文献

(1) 中村・大庭・渡部・石塚・藤本・堀・マクダーモット・南：“音声認識システムSOLONの日本語話し言葉コーパスによる評価（2006年版），”情処学研報，2006-SLP-64, pp.251-256, 2006.
(2) 牧野・荒木・向井・澤田：“ブラインドな処理が可能な音源分離技術，”NTT技術ジャーナル，Vol.15，No.12，pp.8-12，2003.

木下　慶介/ 中谷　智広/ 三好　正人

いつでも，どこでも使える音声インタフェースの実現，それが私たちの夢です．

◆問い合わせ先
NTTコミュニケーション科学基礎研究所
メディア情報研究グループ
TEL　0774-93-5322
FAX　0774-93-5158
E-mail　kinoshita@cslab.kecl.ntt.co.jp