雑音に強い音声認識システム

効率的かつ効果的な雑音補償で高精度に音声を認識

音声認識とは機械が人の話し声を理解できるようにする技術です。近年，静かな環境で非常に高い認識性能を持つ音声認識システムが構築されていますが，騒音が大きい環境ではしばしば性能が大きく劣化してしまいます。

東芝はこれまで，音声認識システムの耐雑音性向上に取り組んできました。今回，雑音問題を効率的かつ効果的に扱う新たな手法PCMLLR（予測制限付き最ゆう線形回帰）を開発し，大幅な音声認識性能の改善を実現しました。VTS（ベクトルテイラー級数）と呼ばれる手法を改良したもので，雑音環境下の米語（米国英語）の音声データベースを用いた評価において，VTSよりも少ない計算量で良い認識性能を示しました。

音声認識の仕組みと課題

人が話したことばをコンピュータに理解させる音声認識は，非常に挑戦的な研究分野として，広く研究開発が行われており，機器の音声コントロールやディクテーション（口述筆記）など，様々な応用が考えられています。

音声認識システムの基本構成を**図1**に示します。入力音声はまず特徴抽出部に入力され，音響特徴量に変換されます。そして，この特徴量がデコーダ部に送られ，音響モデルとのマッチングによって発話内容を表すテキストが求められます。音響モデルは，多数の確率分布を用いて音声の統計的な性質をモデル化しており，あらかじめ大量の学習データを用いて構築されています。

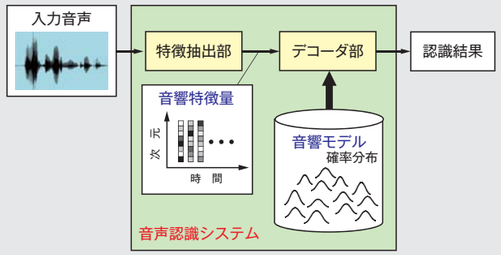

図1．音声認識システムの基本構成 ― 入力された音声から特徴抽出部で特徴量を求め，デコーダ部で音響モデルとマッチングして発話内容を表すテキストが求められます。

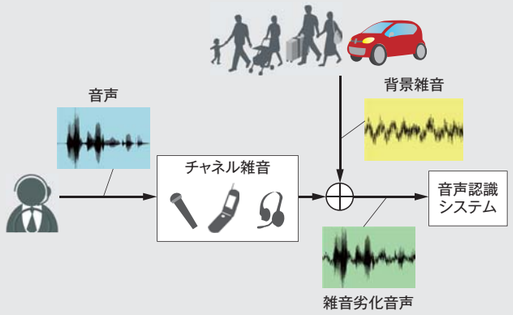

図2．入力音声に影響を与える雑音 ― 自動車の音や他人の話し声などの背景雑音だけではなく，録音機器の違いなどによるチャネル雑音の影響も受けます。

音声認識の課題の一つに，雑音対策が挙げられます。音声は，自動車の音や他人の話し声などの背景雑音だけではなく，録音機器の違いなどによるチャネル雑音の影響も受けます（**図2**）。これらの雑音の影響により，入力音声と音響モデルとの間にミスマッチが生じ，認識性能が低下します。

この問題に対し，これまで提案された多くの解決手法は，入力音声から雑音を除去する特徴量補償と，音響モデルを雑音に適応させるモデル補償に大別されます。モデル補償は一般に特徴量補償よりも良い性能が得られる反面，計算コストが大きいという問題があります。

これに対し東芝は，計算コストを削減した新たなモデル補償の手法を開発しました。

効率的で効果的なモデル補償

新たなモデル補償は，これまで提案された中でもっとも性能の良い手法の一つであるVTSに基づいています。

雑音劣化音声の特徴量（y）は，クリーン音声の特徴量（x），背景雑音の特徴量（n），チャネル雑音の特徴量（h）が与えられたとき，次の雑音劣化モデルで与えられます。

$$y = C\ln(\exp(C^{-1}(x+h)) + \exp(C^{-1}n))$$

C：コサイン変換行列

この非線形な雑音劣化モデルを線形化するため，1次のテイラー展開を適用します。**図3**に示すように，このテイラー展開は音響モデルの確率分布一つひとつに適用され，モデル補償もそれぞれ別々に行われます。そのため，VTSでは非常に高い計算コストが掛かります。

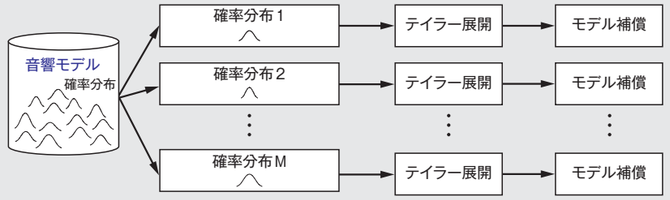

図3．VTSの手続き ― それぞれの確率分布に対してテイラー展開とモデル補償を行います。

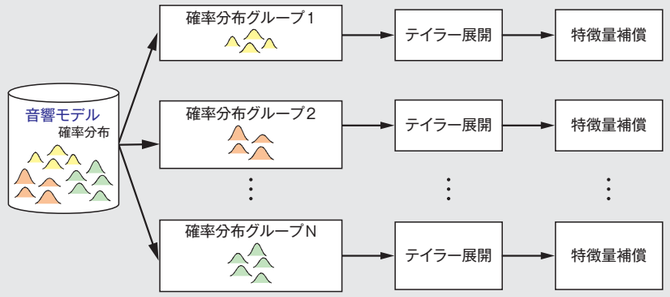

図4．PCMLLRの手続き ― 確率分布をグループに分類し，グループごとにテイラー展開と特徴量補償を行います。

表1．音声認識の正解率

	適用手法		
	雑音補償なし	VTS	PCMLLR
音声認識の正解率（%）	56.72	91.47	92.72

これに対し当社は，計算コストを削減するため，PCMLLRと呼ぶ新たな手法を開発しました。

PCMLLRでは**図4**に示すように，まず，確率分布間の類似度に基づいて音響モデルの確率分布を複数のグループに分類します。そして，一つのグループに対して1回だけテイラー展開を行い，同一グループに属する確率分布でこのテイラー展開を共有します。グループの数は確率分布の数と比べて非常に少ないため，テイラー展開の数を大幅に削減することができます。

次に，グループごとのテイラー展開に基づいて確率分布の補償を行います。このとき，一つひとつの確率分布に対して独立にモデル補償を行う代わりに，それを近似するグループごとの特徴量補償を行います。特徴量補償では，モデル補償で得られる確率分布と特徴量補償で得られる確率分布との距離を最小化する線形変換を求め，特徴量に適用します。

このようにPCMLLRは，VTSでもっとも計算コストの掛かる部分をグループ化し特徴量補償で近似するため，はるかに高速に動作します。音響モデルに含まれるM個の確率分布をN個のグループに分類する場合，PCMLLRではVTSに比べて計算コストをN/M倍に削減できます。典型的なN/Mの値は1/100以下であり，PCMLLRにより大幅に計算コストを削減できます。更に，PCMLLRではVTSでは考慮されていない特徴量の次元間の相関もモデル化できることから，PCMLLRの認識性能がVTSを上回る可能性も考えられます。

音声認識性能の評価

一般に広く用いられている米語の音声データベースを用いて，VTSとPCMLLRの音声認識性能を評価しました。この音声は，背景雑音とチャネル雑音の両方を含んでいます。

音声認識の正解率は**表1**に示すように，雑音補償を行わない場合が56.72％であるのに対して，VTSを用いると91.47％に向上し，PCMLLRでは更に高い92.72％の正解率を得ることができました。

このPCMLLRの正解率は，これまで発表された，同一の米語の音声データベースを用いた評価の中でもっとも良いものの一つです。PCMLLRがVTSよりも何倍も高速であることを考えると，PCMLLRは，雑音対策における最良の選択肢であると考えられます。

今後の展望

音声認識が雑音の影響を強く受けることや，PCMLLRの効率と性能の高さを考えると，PCMLLRは，今後の音声認識アプリケーションにとって重要な技術の一つであると考えられます。

これまでPCMLLRでは雑音だけを扱ってきましたが，今後は，話者の違いなど雑音以外の変動に対する補償も検討する予定です。

許　海天（Haitian Xu）
元 東芝欧州研究所
ケンブリッジ研究所リサーチエンジニア

益子　貴史
研究開発センター
知識メディアラボラトリー主任研究員