不特定話者連続音声認識

“ すぐ使えて，抜群の性能 ”が拓く多彩な応用

ことばを使うのは人類だけです。一方，日本語のかな漢字書きはわれわれの祖先の大発明であり，われわれへの贈り物です。この間を橋渡しする音声ワープロ（あるいは音声による文章入力ソフトウェア）の実現は長年の夢でした。当社は，ユーザーの声の登録を必要としない，文章入力ソフトウェア“ 東芝音声システムv5 ”を,1999年にパソコン（PC）“ Dyna Book ”にプリ インストールして製品化しました。更に，2000年11月には，パッケージソフトウェアとして“ LaLa Voice$_{TM}$ 2001 ”を商品化しました。

連続音声認識とは，単語ごとにくぎって話すなどの制約を付けず，普通の話し方のことばを認識する技術です。不特定話者認識ではユーザーの声の事前登録を必要とせず，だれの声でも認識します。音声による文章入力を多くの人が真に使えるようにするためには，この二つが必須です。

図1．連続音声認識の幅広い応用　音声情報を扱う幅広い応用へ展開していきます。

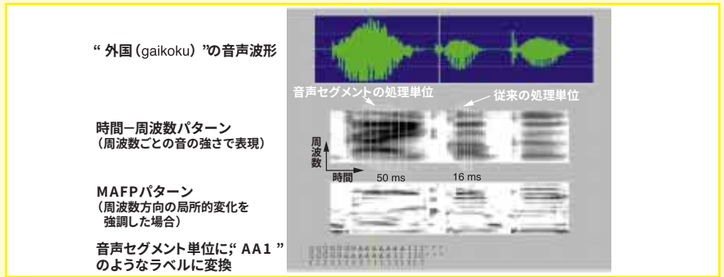

図2．音声認識のための特徴抽出　音声セグメントの処理単位をラベル変換して，日本語言語情報を背景に，音味のある音として扱っています。

■声の事前登録の常識を打ち破る

従来，連続音声認識のためには，声の事前登録が常識とされてきました。当社は事前登録なしで，しかもトップクラスの性能を達成しています。

声の登録を完全になしにすることで，以下のようなメリットがあります。

⑴　登録の手間がなく，すぐ使える。
⑵　周囲環境や声が変っても認識する。
⑶　誤登録による性能低下がない。

また，次のように応用が広がります。

⑴　ほかの人の録音データを認識して，会議録・講演録の自動作成
⑵　テレビ放送のアナウンサーの声を認識して，字幕作成やキーワードの抽出
⑶　家族みんなによる，ホームコントロールや情報入力
⑷　あふれる音声の自動文字化による，デジタルアーカイブの作成
⑸　不特定多数が使う券売機など，ソシアルシステムへの適用
⑹　使用環境の変動の大きいモバイル機器の操作や情報入力
⑺　翻訳ソフトウェアでのチャット（PC間でのメッセージのやり取り）や，通訳機でのコミュニケーション
⑻　電話での予約や情報検索
⑼　ロボットとの対話や命令

このように，当社の技術は，文章入力のために発声する応用にとどまらず，既に音声となっている情報を扱う応用へと大きく展開できます（**図1**）。

■音をつながりで見る

声は人によって違うのはもちろん，言葉の前後の関係，周囲の環境，話す人のその日，そのときの体調や発声ごとでも変化します。

音声の特徴抽出法を**図2**に説明します。一般には，入力音声を16 ms程度の短い区間に分けて，その区間の音声波形データから特徴量を計算し，これを直接，認識に使用します。一方，当社では，日本語の様々な音の特徴を表現する50 ms程度の“ 音声セグメント ”単位を提案しました。“ 音声セグメント ”では音の前後関係を含めて照合できるという特長がありますが（**図3**），データ量が多くなるため“ 音声セグメント ”を示すラベル（名前）に変換しています。すなわち，音声波形を単なるこま切れの信号として扱うのではなく，日本語言語情報を背景にした意味のある音として扱っています。

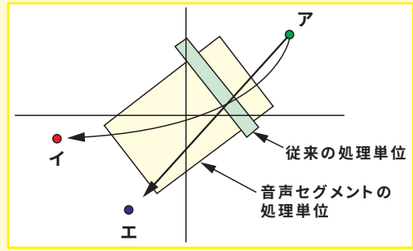

図3．音声セグメントと従来の処理単位の違い　“ ア ”から“ イ ”と，“ ア ”から“ エ ”への変化では，途中のルートが似ているが，50 ms程度の広い範囲を扱う“ 音声セグメント ”なら確実に区別することができます。

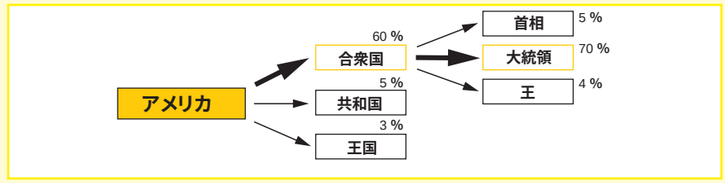

図4．単語のつながりの概念　例えば，“ アメリカ ”の後には“ 合衆国 ”，“ 共和国 ”，“ 王国 ”の中だと，“ 合衆国 ”がくることが多い。このように，あとにきやすい単語をあらかじめ想定しておくことで，より確実な認識が可能になります。

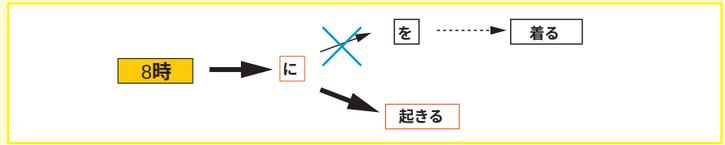

図5．文法的接続チェックの概念　普通の日本語では，“ に ”の後に“ を ”がつながることはない。したがって，同じ発音の“ おきる ”でも，正しく“ 起きる ”が選択されます。

■音の変化をとらえる

人は音声の絶対的なレベル（強弱）だけでなく，その変化をも感じていると考えられます。その変化を特徴量として抽出するために，“ 複合音響特徴平面；MAFP（Multiple Acoustic Feature Plane）”と呼ぶ，音の細かい変化を扱うことのできる方法を開発しました（図2）。これによって，音声の局所的な強弱だけでなく，音の周波数方向と時間方向の変化を含めた，二次元の特徴量として音声を扱えます。MAFPでは，従来方法に対して，誤り率が1/2近くになるという実験結果も得られています。

■単語のつながりを見る

連続音声認識では，一単語だけで判断するのではなく，例えば，後に続くことばを想定するなど，日本語の単語の並び方の常識を利用しています（**図4**）。

また，当社ではかな漢字変換，機械翻訳などで培った日本語処理技術をフルに活用して，精度の高い“ 形態素解析・言語処理技術；JUMP（Japanese Analyzing Unit with Morphological Procedure）”を開発しました。これは，文法的な接続関係をチェックして，日本語として正しくない文は出さない技術です（**図5**）。ここでは，同音異義語の選択も行います。

更に，JUMPでは大量の文章データから単語のほか，助詞なども含めた形態素を自動的に抽出して，そのつながり方を統計的に調べておきます。大量のデータを扱うためには，この自動的に行う形態素解析の精度が認識性能に大きく影響します。

したがって，JUMPを活用して文章入力ソフトウェアを専門分野，例えば，X線CT（Computed Tomography），MRI（Magnetic Resonance Imaging）などから得られる医用画像の診断レポート向けに適応できます。当社で収集した評価データを用いた実験によれば，適応した医用文章入力ソフトウェアでの認識精度は98 %と，十分に実用的です。

■コミュニケーション文化が変る

当社は，78年のかな漢字変換技術の開発とワープロの発売によって，日本の文字文化に大きな影響を与えました。今後，当社の音声認識（不特定話者連続音声認識）技術は，音声情報を扱う幅広い応用を通して，新しいコミュニケーション時代に貢献していきます。

研究開発センター
マルチメディアラボラトリー主任研究員
松浦　博