特集

雑音環境下における音声認識技術

Technologies for Speech Recognition in Noisy Environment

あらまし

音声認識を実際の環境で使用する際に雑音が大きな問題となる。本稿では，雑音混じりの音声から，発話した内容を認識する技術について述べる。まず，雑音対策を行わない音声認識システムで，雑音の影響による認識性能劣化の状況から，雑音対策の必要性を示した後に，雑音対策のおおまかな分類と実現方法の例について述べる。さらに，不特定話者単語音声認識システムの雑音対策として，推定した雑音パターンを雑音混じりの音声から差し引く方法を紹介し，実際に雑音下で発話された大量の音声データを用いた単語音声認識実験を行って，雑音対策が認識性能の向上に効果があることを示す。

Abstract

Noise is a big obstacle to achieving speech recognition in a practical environment. This paper describes speech recognition technologies which extract text from noisy speech.

First, we explain how recognition systems without countermeasures for noise have a poor recognition performance for noisy speech. Next, we introduce some techniques for noisy speech recognition. Finally, we describe a speaker-independent word recognition system based on a spectrum subtraction technique and present some experimental results of speech recognition in a noisy environment.

岩見田　均（いわみだ　ひとし）

1983年北海道大学工学部電子工学科卒。同年（株）富士通研究所入社。以来音声認識の研究に従事。1988年～1991年（株）ATR 視聴覚機構研究所に出向。1991年日本音響学会粟屋潔学術奨励賞受賞。
パーソナルシステム研究所ヒューマンインタフェース研究部

木村晋太（きむら　しんた）

1980年神戸大学大学院工学研究科修士課程了。同年（株）富士通研究所入社。以来音声・音響処理技術の研究に従事。
パーソナルシステム研究所ヒューマンインタフェース研究部

まえがき

近年，音声認識を用いた製品の提供が盛んになりつつある。現状の音声認識技術では，はっきりと発話する，静かな環境で使う，などの様々な制約が課せられないと，良い性能が得られないという問題がある。これらの制約は，音声認識の普及を妨げる大きな要因になっている。

本稿では，音声認識を実際に使用する際に大きな問題となる雑音に注目して，雑音混じりの音声から，発話した内容を認識する技術について述べる。

以下，まず，雑音対策の必要性，雑音対策の分類について述べ，つぎに，雑音下の音声認識システムと，それを用いた認識実験結果について述べる。

雑音環境下の音声

雑音には様々な種類のものがあるが，本稿では，空調の音のように，マイクロホンに定常的に付加される雑音を対象とする。

雑音が付加された音声波形の例を 図-1 に示す。これらはすべて「あさひ」という発話の音声波形に対して，通常の音声認識処理の前処理として行う，周波数の高域を強調する処理（高域強調）を施した後の音声波形である。(a) が最も雑音レベルが高く，(c) が最も低い。(b)，(c) では，音声の部分の波形を確認することができるが，(a) では雑音に埋もれて，音声部分の波形の確認は困難である。音声に対してどれくらいの雑音が付加されているかを示す指標として音声対雑音比（以下，SN 比）がよく用いられる。通常，音声のパワー (S) と雑音のパワー (N) の比を SN 比と呼ぶが，雑音下で発話された音声の場合，音声のみのパワーを算出するのは困難なので，本稿では，音声＋雑音のパワー（S＋N）と雑音のパワー (N) の比を SN 比と定義する。図-1 のデータの SN 比は，(a) が 8 dB，(b) が 16 dB，(c) が 27 dB である。

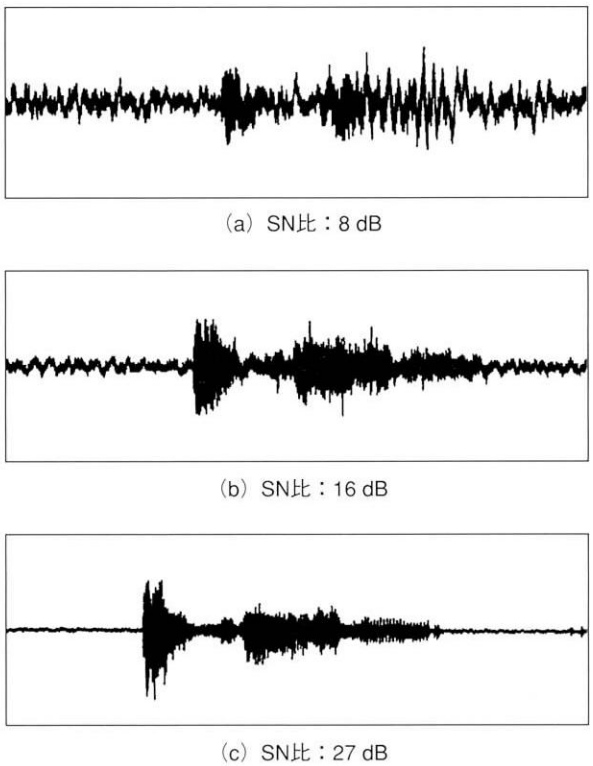


図-1　雑音下の音声波形の例
Fig.1-Speech samples under noisy environment.

雑音対策の必要性

雑音を考慮しない音声認識システムを用いて，雑音下の音声を認識した場合の，SN 比と 100 単語認識率の関係を図-2 に示す。この結果から，SN 比が小さくなるに従って（雑音が大きくなるに従って），認識率が急激に低下しているのが分る。

なお，図-2 において最も SN 比が小さい音声でも，人間が聴けばほとんどが正確に認識できる程度のものである。人間の聴覚が，雑音に対して柔軟に対応できているのに対し，これまでの自動音声認識技術は雑音に非常に弱く，雑音対策の必要性があるといえる。

雑音対策の分類

典型的な音声認識処理の流れを 図-3 に示す。雑音対策

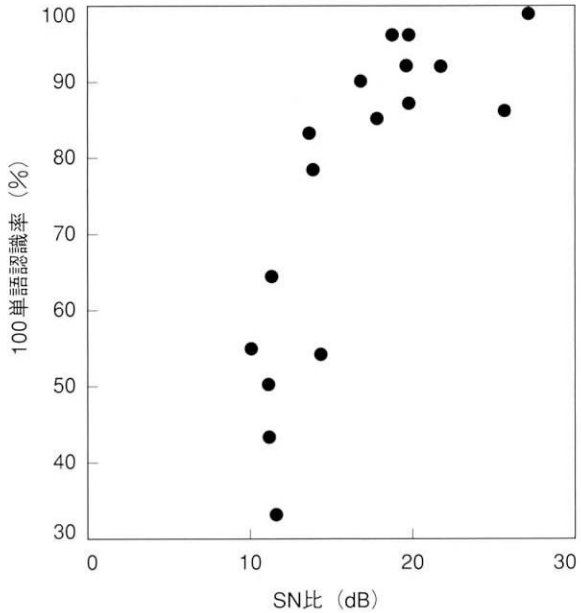


図-2　SN 比と認識率の関係
Fig.2-Recognition performance in noise.

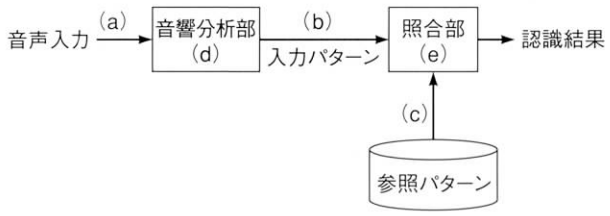


図-3　音声認識処理の流れ
Fig.3-System configuration of speech recognition.

としては，図中の (a)～(c) の各位置に，雑音対策処理を追加する方法と，(d)，(e) の処理を雑音対策用に変更する方法が考えられる。それぞれの方法における雑音対策の考え方と例について以下で述べる。

● 入力音声波形に対するもの

図中 (a) の位置で，音声波形中の雑音成分を少なくする，あるいは，音声成分を強調する方法であり，複数マイクロホンを使うものが一般的である。[1]

● 入力パターンに対するもの

図中 (b) の位置で，雑音の特徴を推定して，入力パターンの特徴から取り除く方法であり，スペクトルサブトラクション法 (以下，SS 法) が代表的である。SS 法は，音声入力の直前のスペクトルパターンを雑音パターンとし，雑音下の入力音声のスペクトルパターンから雑音パターンを差し引くものである。SS 法は，処理量が少なく実現も比較的容易であるため，広く用いられている。

● 参照パターンに対するもの

雑音の特徴を推定して，図中 (c) の位置で参照パターンに加えてから，入力パターンと照合する方法であり，スペクトルアディション法が代表的である。前記の入力パターンに対するものが，雑音を除いた音声同士を照合するのに対し，この方法は，雑音混じり音声同士を照合することになる。また，前記の方法は，入力パターン一つに対して処理を行うだけで済むのに対し，この方法は，複数の参照パターンに対して処理をしなければならないので，処理量が比較的多くなる。

● 分析方法によるもの

雑音が含まれる場合でも，雑音がない場合に近い分析結果が得られるような分析を行うものである。雑音の周波数成分が偏っている場合に，その付近の周波数成分を弱くするフィルタを前処理として用いる方法や，定常的な雑音の影響を受けにくくするために，スペクトルの時間変化をとらえる分析方法などがある。

● 照合方法によるもの

照合時に雑音の影響を低減するもので，入力パターンと参照パターンとの距離を算出する際に，プロジェクション距離を用いる方法，入力音声の先頭や最後の部分が欠落することに対処して，参照パターンの先頭や最後の部分を欠落させたパターンを追加した複数の参照パターンを持つ方法などがある。

雑音環境下における音声認識

● 評価用音声データ

評価用の音声データとして，実際の雑音環境下で収録した音声を用いる。音声は自動車内で話者 4 名が，日本国内の地名 100 単語を 2 回発話したもので，データの総量は，4 名 × 2 回 × 100 単語 = 800 単語である。

● 認識方式

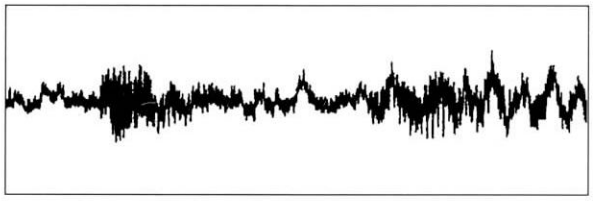
認識の基本方式として，音響セグメントネットワークによる音声認識方式[2] を用いる。この方式では，認識したい語彙のリストは，文字列として与えればよいので，大語彙認識や，頻繁に認識したい語彙を変更しながらの認識に向いている。上述した参照パターンとしては，日本語の音素体系に基づいて定義した 28 種類の発音のスペクトル (音響テンプレート) を用いる。音響テンプレートは，あらかじめ多数話者の大量の音声から抽出して作成しておく。これにより，評価用データはだれが発話したものでもかまわない，不特定話者音声認識を行うことができる。

● ハイパスフィルタ

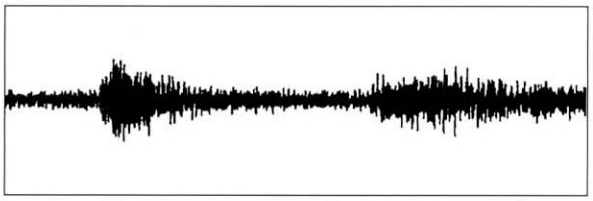
自動車内収録音声データに含まれる雑音成分は，周波数の低い成分が比較的強いので，その影響を少なくするために，カットオフ周波数が 300 Hz のハイパスフィルタ (以下，HPF) を用いる。これにより，300 Hz 以下の成分を低減することができる。HPF をかける前の音声波形と，かけた後の音声波形の例を 図-4 に示す。この図から，HPF をかけた後の方が，音声部分の波形が比較的確認しやすいことが分る。

● スペクトルサブトラクション

前章で述べた SS 法を用いる。雑音のスペクトルは，認識する音声に先行する非音声区間の複数フレーム (15 ミリ秒間隔のタイムスライス) のスペクトルパターンを平均して求める。これを，認識に用いる音声区間の各フレームのスペクトルから差し引くことで，スペクトル中の雑音成分を除去する。SS 法の実施の様子を 図-5 に示す。(a) は元の音声のスペクトル，(b) は雑音が加えられた音声のスペクトル，(c) は推定した雑音スペクトル，(d) は SS 実施後のスペクトルである。雑音のスペクトルが強い低域において，スペクトルが減算され，(a) の元の音声のスペクトルに近くなっている様子が分る。

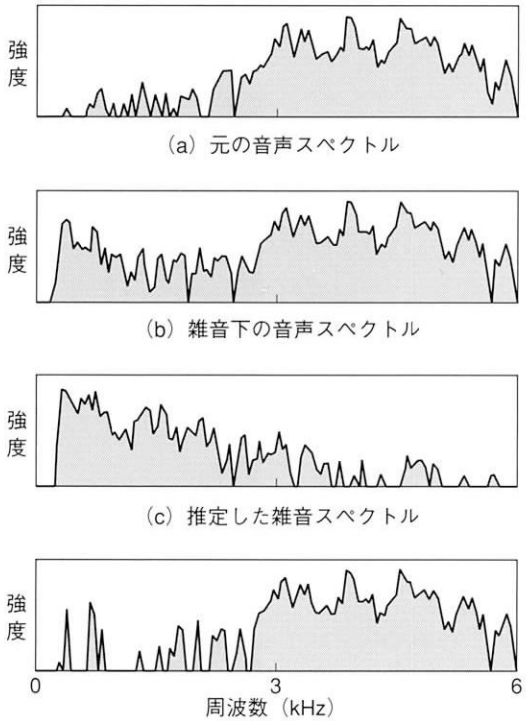


(a) HPFをかける前



(b) HPFをかけた後

図-4　ハイパスフィルタをかける前後の音声波形
Fig.4-Speech samples before and after high pass filtering.



(a) 元の音声スペクトル
(b) 雑音下の音声スペクトル
(c) 推定した雑音スペクトル
(d) 処理後の音声スペクトル

図-5　スペクトルサブトラクション（SS法）実施の様子
Fig.5-Aspect of a spectrum subtraction.

なお，この例の音声は「き」の子音部である。

● 認識実験の結果

認識実験の結果を図-6に示す。認識語彙は地名100単語である。上記雑音対策を全く行わない場合の結果も併せて示す。この結果から，上述した雑音対策を行うことで，認識性能が向上することが分る。とくに，SN比が低い場合に，大幅に認識率が向上しており，SN比が良くない

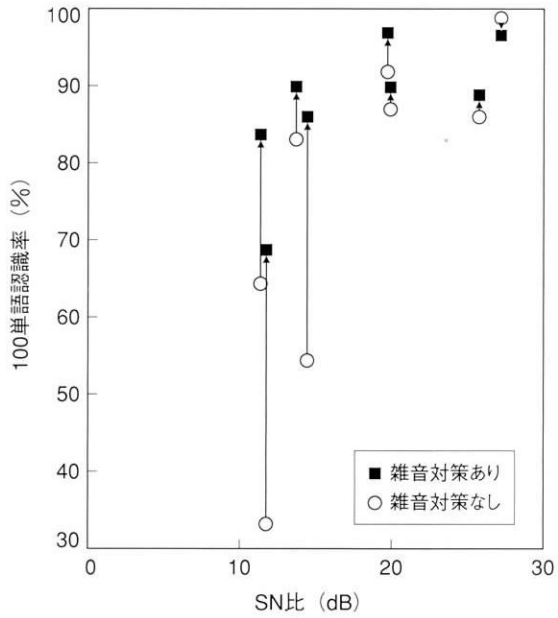


図-6　雑音対策と100単語認識率
Fig.6-Recognition performance with noise robust techniques.

音声データに対しては，雑音対策が不可欠であるということがいえる。

今後の研究課題

今回の実験では用いていない，音声波形そのものに対する雑音対策を組み合わせた場合の評価と，今回は対象としなかった，非定常的な雑音に対する検討を行う必要がある。非定常雑音の対策としては，参照パターンに雑音パターンを追加して，非定常雑音が加わった音声をモデル化する方法や，音声の前後の雑音を照合から除くためのワードスポッティング法などが考えられる。

む　す　び

音声認識を実際に使用する際に大きな問題となる雑音に注目し，雑音対策の必要性と雑音対策例を紹介した。そして，実際に雑音下で発話された音声データを使った認識実験を行なって，雑音対策の効果を示した。とくに，雑音の割合が多い音声データに対しては，雑音対策を行うことで，大幅に認識率が向上することを確認した。

参考文献

(1) 松尾ほか：マイクロホンアレイを用いた音声入力インタフェース．*FUJITSU*，**49**，1，pp.80-84（1998）.
(2) 木村：音響セグメントネットワークを用いた大語い音声認識．信学論（D-II），J77D-II，3，pp.475-482（1994）.