NTT音声認識技術の最前線
–認識エンジンと音声区間検出を中心に–

中村 篤

NTT コミュニケーション科学基礎研究所
日本電信電話株式会社

http://www.kecl.ntt.co.jp/icl/signal/nakamura/

音声認識基礎技術

音声コミュニケーションシーンを「ことば」に書き下す技術

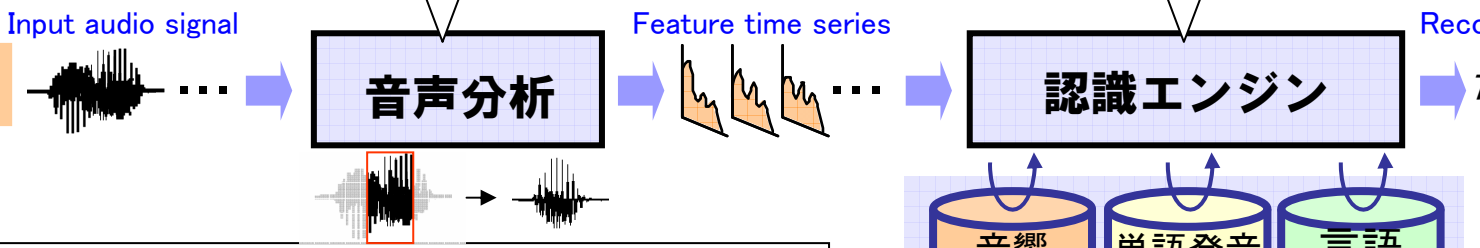
信号処理＋統計的時系列解析

情報構造＋計算モデル・アルゴリズム

- 日常的背景雑音を含む音信号からの頑健な音声区間検出+雑音抑圧(●)とスペクトル特徴量の抽出(●)

- 重み付き有限状態トランスデューサに基づく高速・高精度探索ｱﾙｺﾞﾘｽﾞﾑ

＊超大語彙音声認識（≈1千万語）(●)



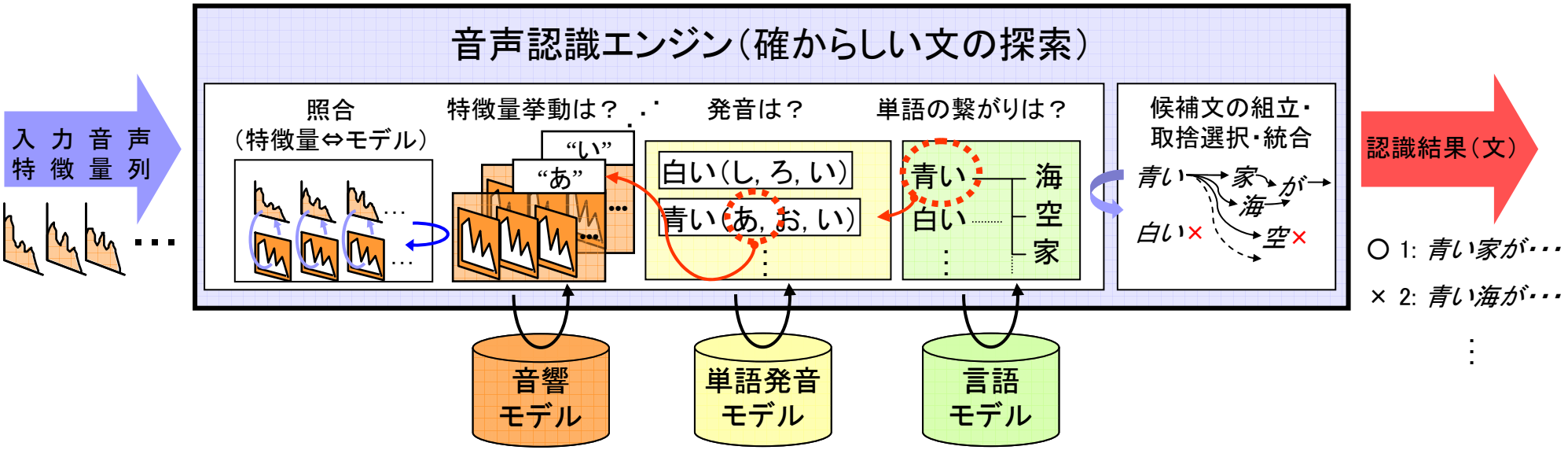
- ベイズ法による音響モデルの設計と音声分類の枠組み(●)
- 音響モデルの識別的学習(●)
- オンライン／オフラインモデル適応（極少量データによるモデル学習）(●)

- 言語モデルの識別的学習(●)
- リッチ・トランスクリプションのための高次言語処理と(●)キーフレーズ抽出(●)

確率・統計モデル＋学習理論応用

世界メディアブラウザ(●)

音声認識の原理



音響モデル:　声の雛形
単語発音モデル:　語の雛形
言語モデル:　文の雛形

} の色々な組合せを試しながら

入力された音に最も合致する**言葉の並びを探しだす**

音声認識エンジンの役割 (1/2)

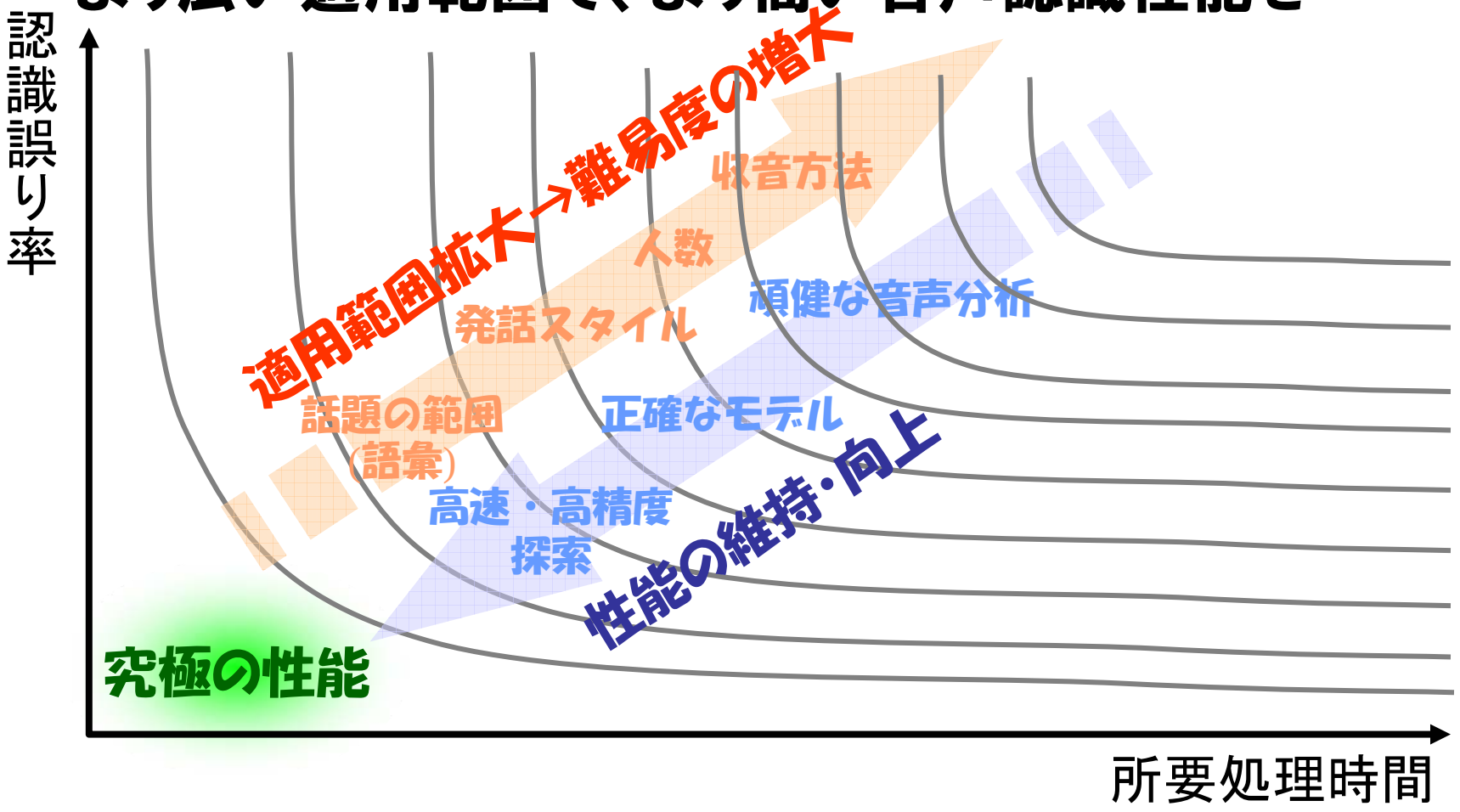
モデルをもとに、単語列の仮説を立て、評価値を計算
最も評価値の高い単語列を効率的に探し出す(「探索」する)

高精度な探索は高精度音声認識の「必要」条件
- 「必要十分」ではない
- ⇒ 完璧な探索のみでは、必ずしも正解単語列を探し出せない
  - ∵地図(モデル)が間違っていては正しい目的地(正解単語列)に辿りつけない
  - ・・・それでも、正確な探索なくして、折角の良いモデルは活かせない

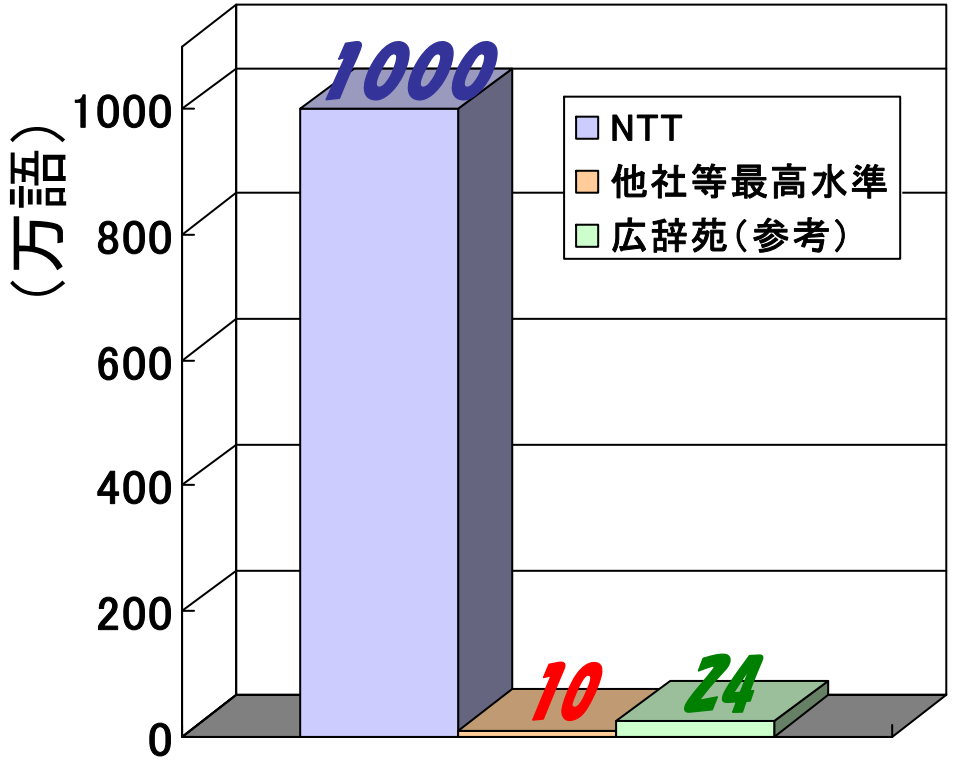
高速な探索は音声認識の用途を広げる強力な武器
- 手っ取り早い高速化は、「近視眼的」探索(e.g.「絞った」ビーム探索)
- ⇒ 安易にスピードを求めても探索精度とトレードオフの関係
  - 慌てて進むと分かれ道で間違えて、思わぬ場所に出てしまう
  - 精度を落とさない「筋の良い」高速探索が理想

音声認識エンジンの役割 (2/2)

**高速・高精度な探索により、音声分析、モデルとともに
より広い適用範囲で、より高い音声認識性能を・・・**



語彙1000万語で実時間動作が可能な超高速連続音声認識メガエンジン



従来方式：高々6～10万語程度

数十万語規模で性能が大きく劣化

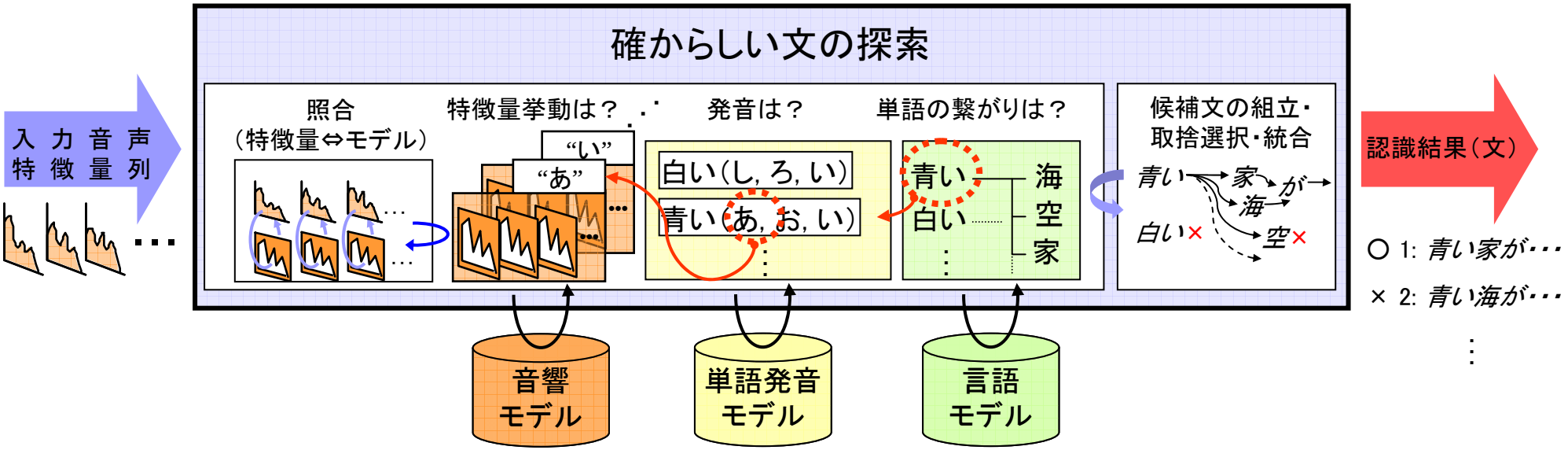
NTT独自の
高速・高精度音声認識技術
“WFST”＋“高速on-the-fly合成法”

超大語彙音声認識：約1千万語
（氏名、地名、組織名、新造語、・・・）

従来不可能であった数百万～1千万語規模の音声認識を独自の高速・高精度音声認識技術により世界で初めて達成

WFST: Weighted Finite-State Transducer（重み付き有限状態トランスデューサ）

従来型(個別モデル型)音声認識



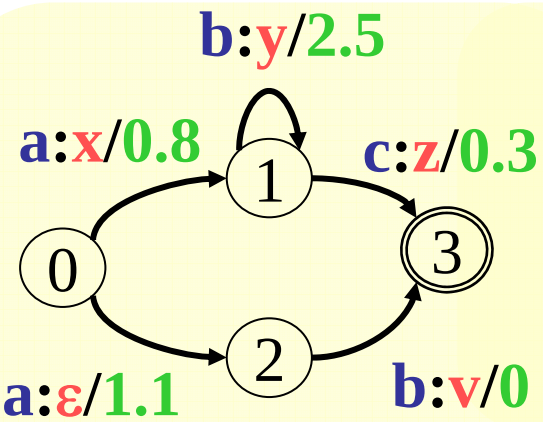
モデルが別々になっているために・・・

モデルのパーツを逐次的に対応付け⇒ 時間がかかる(速度低下)

必要な知識全体を十分見渡せない ⇒ 間違いやすい(精度劣化)

重み付き有限状態トランスデューサ(WFST)によるモデル統合へ

重み付き有限状態トランスデューサ
WFST: Weighted Finite-state Transducer



（入力:出力/重み）

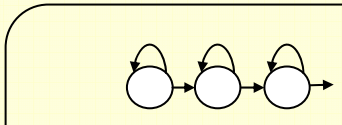
WFSTとは・・・

- オートマトンの一種（状態遷移）
- 入出力シンボル・重みスコア
- 情報変換（Transduction）の汎用計算モデル

a b c ⇒ 変換 ⇒ x y z/3.6

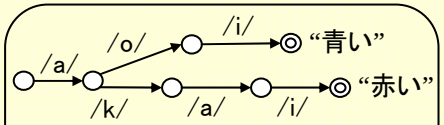
WFSTとして解釈可能

各々が上位言語階層のシンボルへの変換を担う



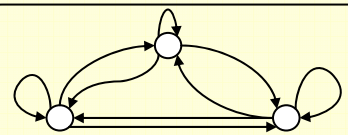
隠れマルコフモデル

特徴量 → 音素



（確率的）正規文法

音素 → 単語



（多重）マルコフモデル

単語 → 文

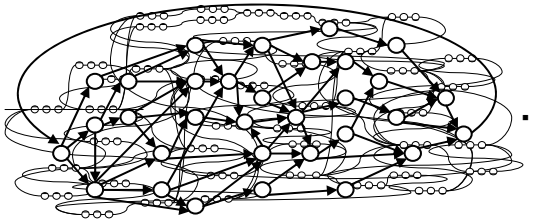
音響モデル

単語発音モデル

言語モデル

合成・最適化

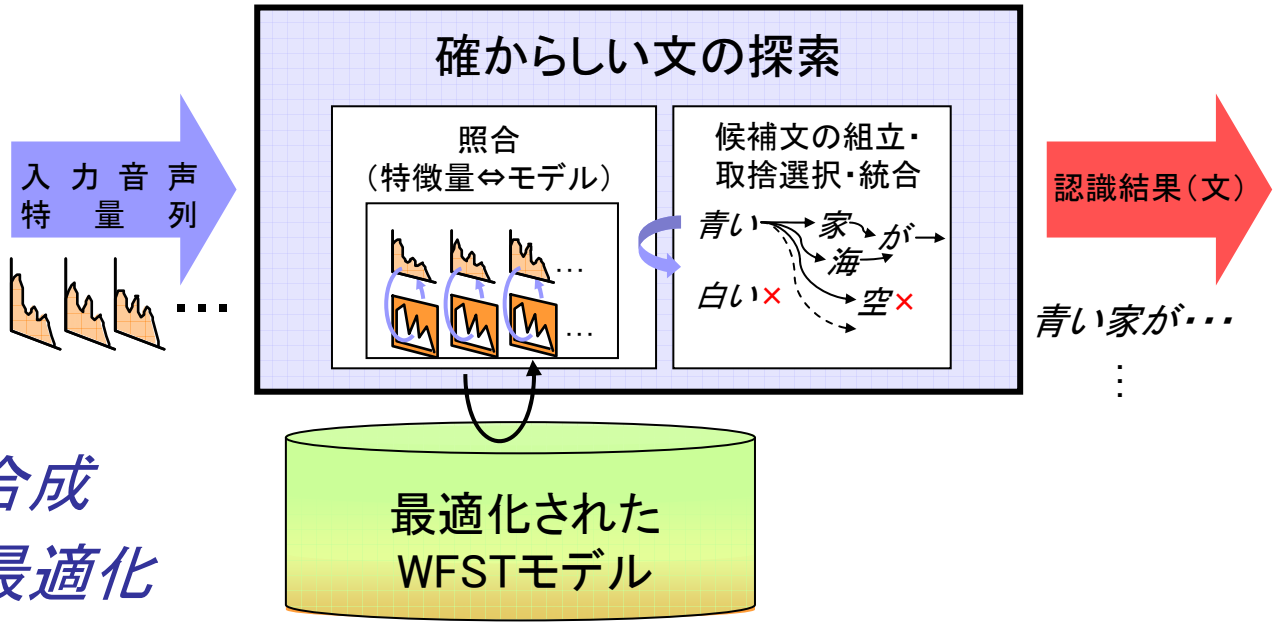
最適化(*)されたWFSTモデル



特徴量 → 文

(*) 等価性を保ったまま「探索向け」の構造に

WFST型音声認識



WFSTの合成

WFSTの最適化

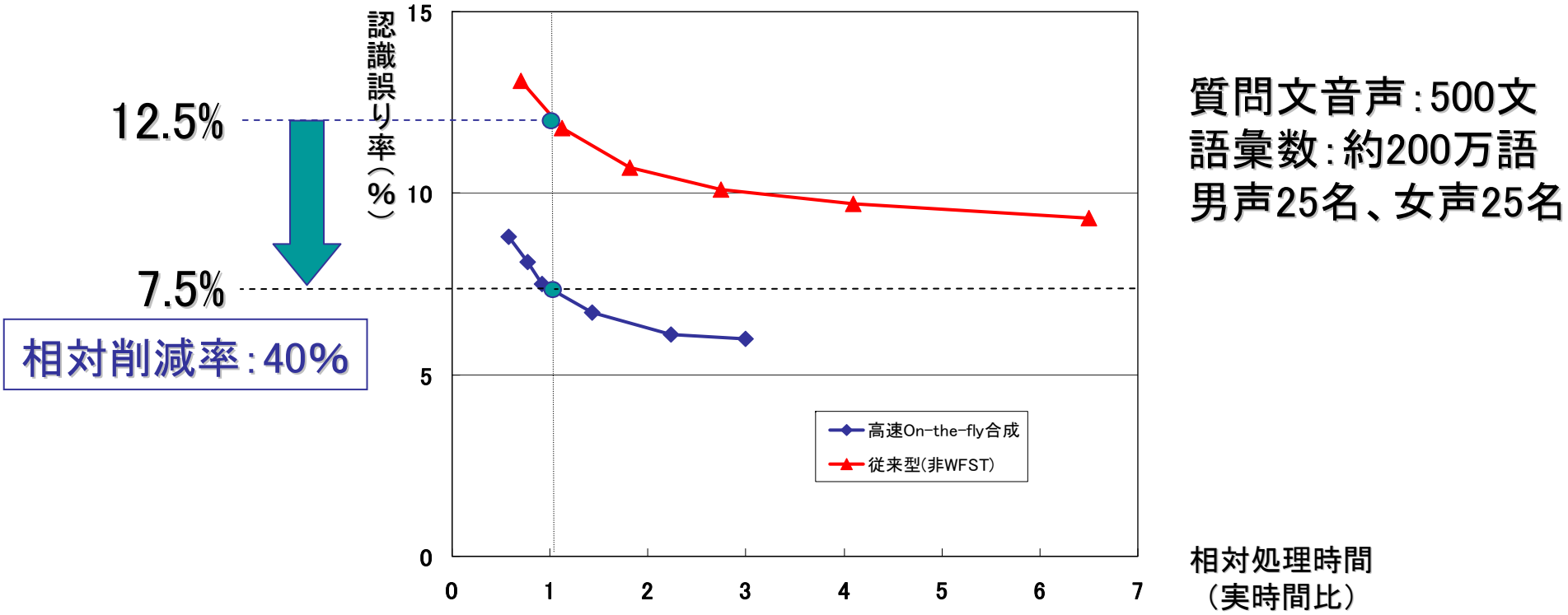
WFSTによって・・・

逐次的対応付けを解消・探索エンジン軽量化 ⇒ 素早く動く

複数モデルを統合・知識源全体を再整理 ⇒ 探しやすい

従来型との性能比較(超大語彙)

速度、精度ともに従来型を凌駕 (約200万語での評価)
1000万語でも同様の傾向



質問文音声:500文
語彙数:約200万語
男声25名、女声25名

高速・高精度探索: 語彙拡大のみならず、多様な難条件に対し普遍的効果

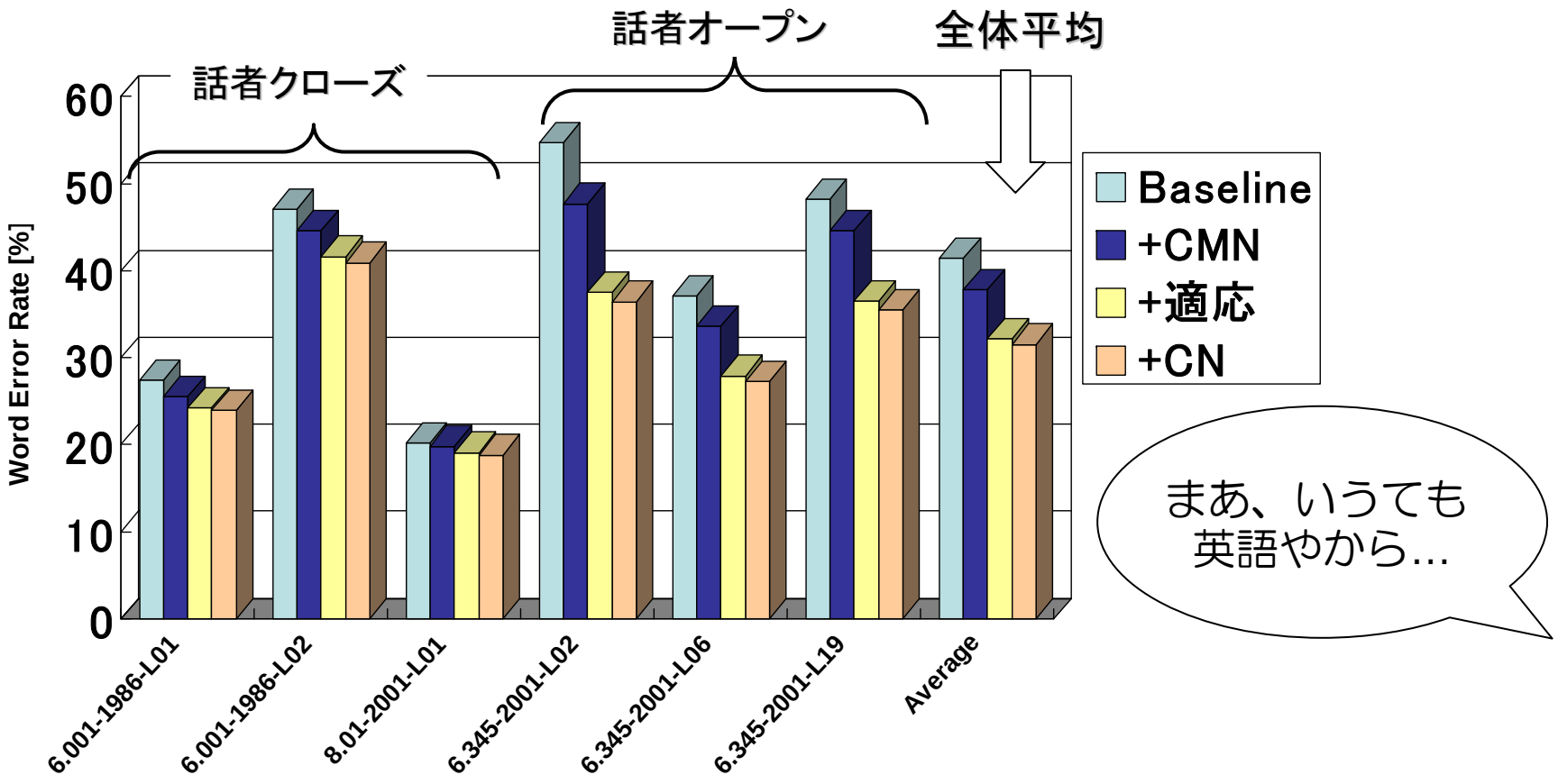
⇒ 実環境で頑健に動作する音声認識の核技術

ビデオ: 英語の講義に字幕をつけてみる

音声トラックをあらかじめ自動認識してテキスト化（無修正）
映像と同期しながらそのまま「カラオケ風」字幕として表示

OCW英語音声認識結果(単語誤り率(%))

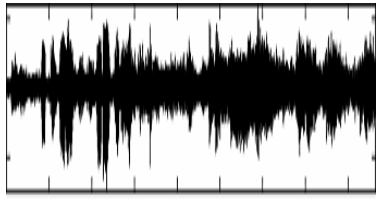
非母国語であることの困難は最小限でそこそこの精度
WFSTにおける抽象化された探索空間最適化が功奏？

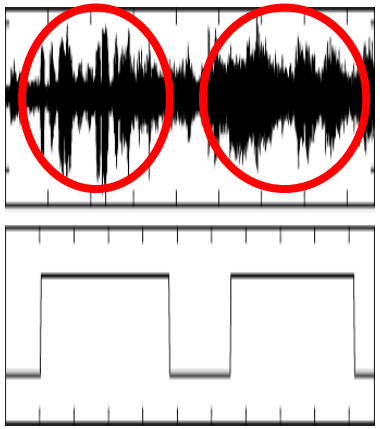


音声区間検出とは

観測信号から目的音声の存在区間を自動検出する技術

観測信号 → **音声区間検出**
VAD:
Voice Activity Detection → 検出結果





⇒ 用途の広い重要技術

音声認識（Hands-free発話 ⇒ 遠隔発話）
携帯電話、Voice over IP（低伝送容量下での無音圧縮）
複数音源の分離、残響除去・制御の基本要素技術
人手によるデータトランスクリプション・アノテーションの前処理、・・・

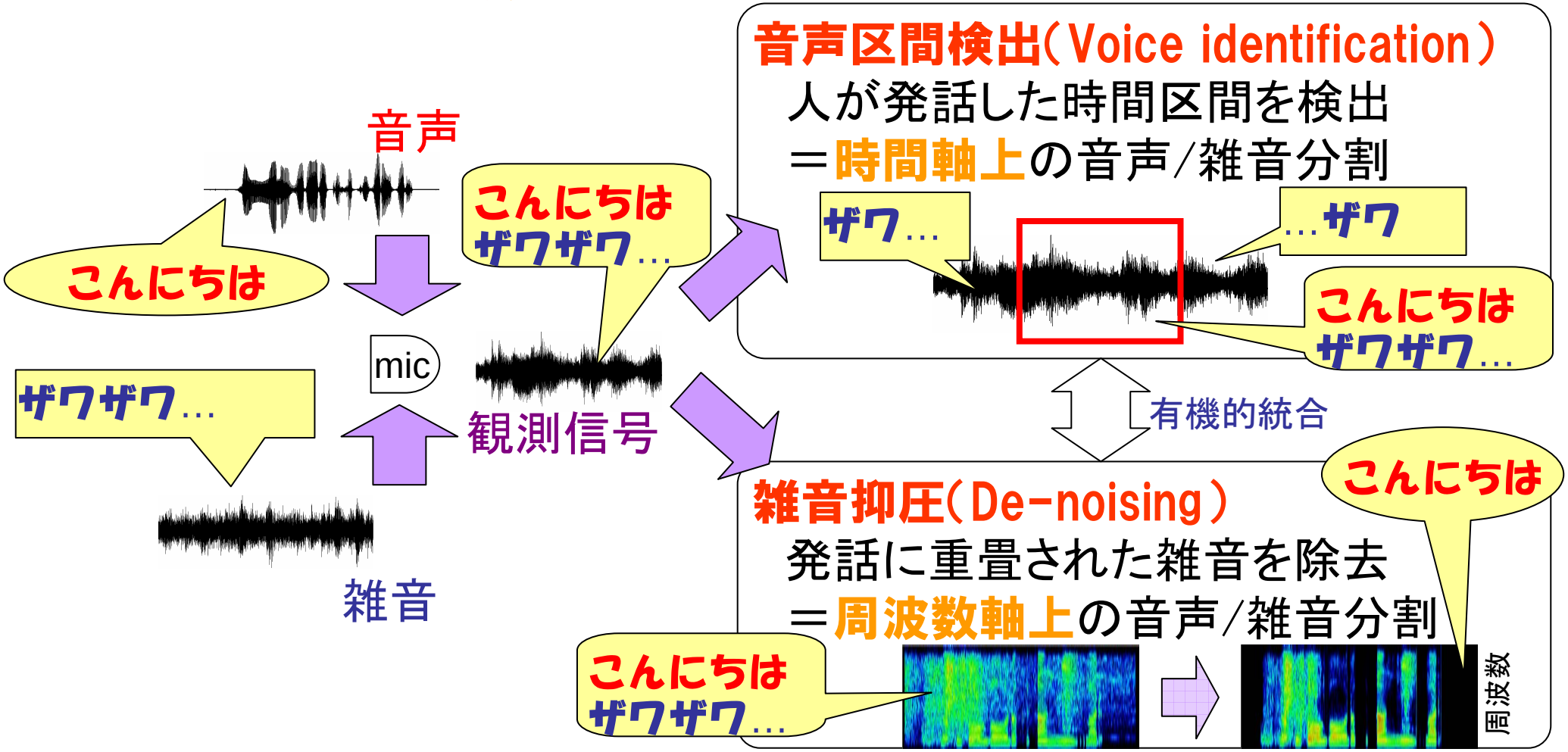
⇒「使える」技術への期待高（というか、期待されて久しい・・・）

雑音下で正確に音声存在区間を検出（重畳雑音も除去してくれればなお良し）
多くの用途に耐えるオンライン高速処理（リアルタイム処理ならばなお良し）

NTTの音声区間検出＋雑音抑圧技術

DIVIDE: Dynamic Integration of Voice Identification and DE-noising

時間・周波数の両軸上で音声と雑音を自動分割



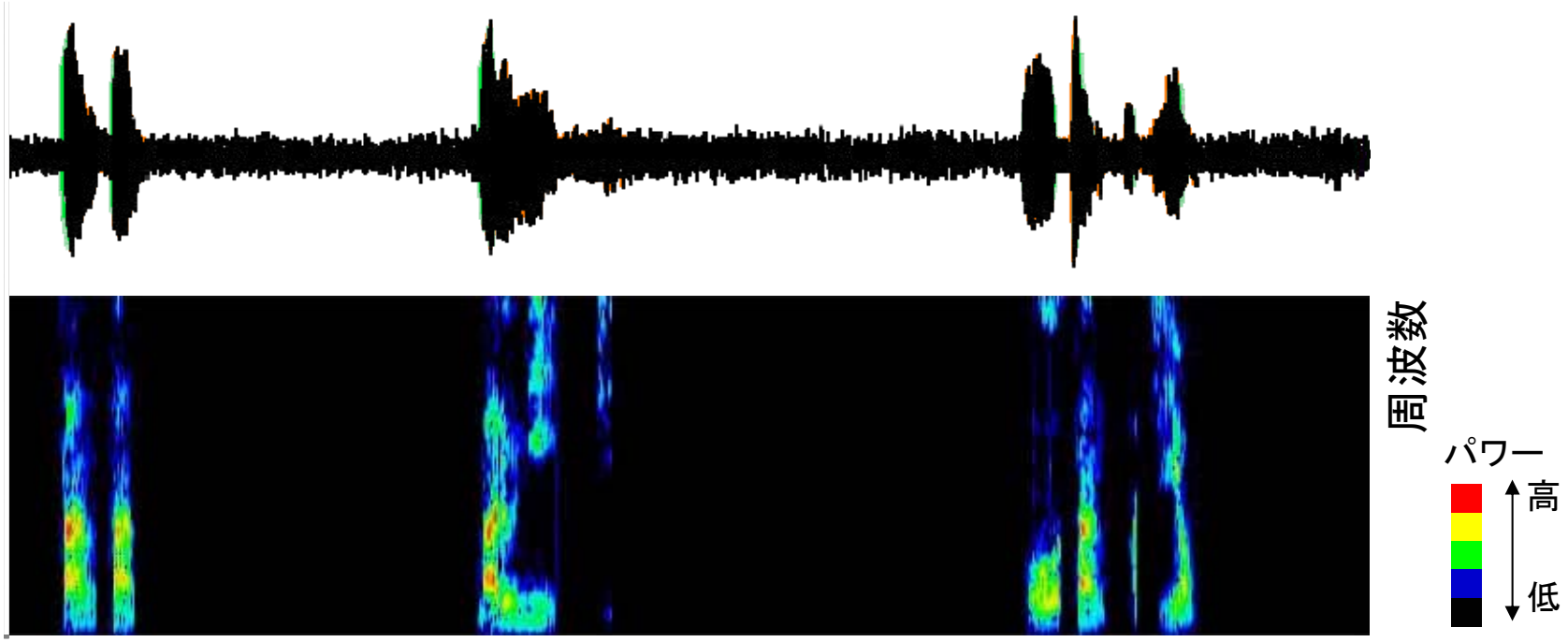
DIVIDEの実行例

正確な音声存在区間の検出と効果的な雑音抑圧を同時にリアルタイムで実現

雑音が重畳された音声（観測信号：9桁数字発声）

音声区間検出結果（DIVIDEのVADのみ適用）← **時間軸上の分割**

音声区間検出＋雑音抑圧結果（DIVIDE適用）← **周波数軸上の分割**

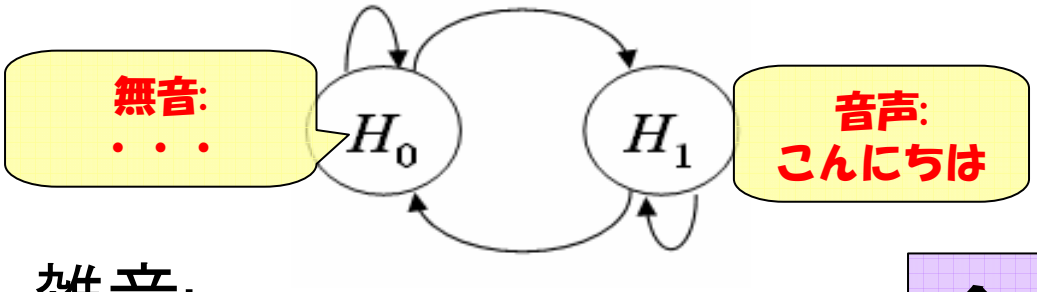


音声区間検出技術の概要

音声と雑音のデータから区間検出モデル、雑音モデルを学習
合成モデル上での最良パス選択による音声/非音声の判定

音声:

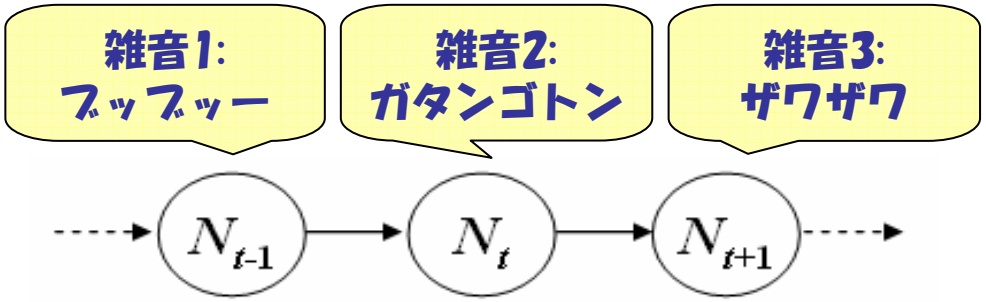
2状態マルコフモデルとして事前学習



雑音:

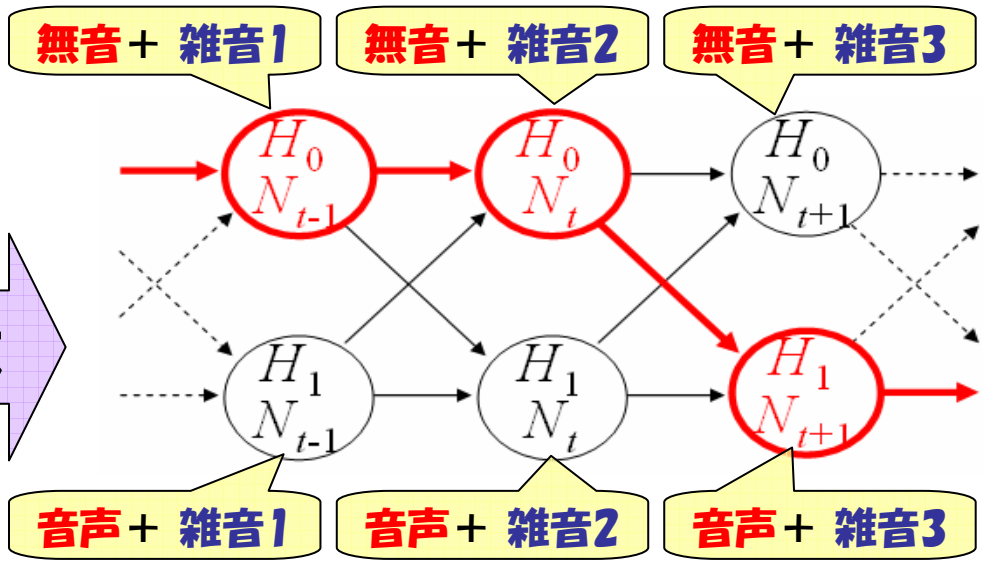
動的追従型モデルとして

時刻フレーム(t)毎にオンライン学習



合成

音声＋雑音: 合成モデル



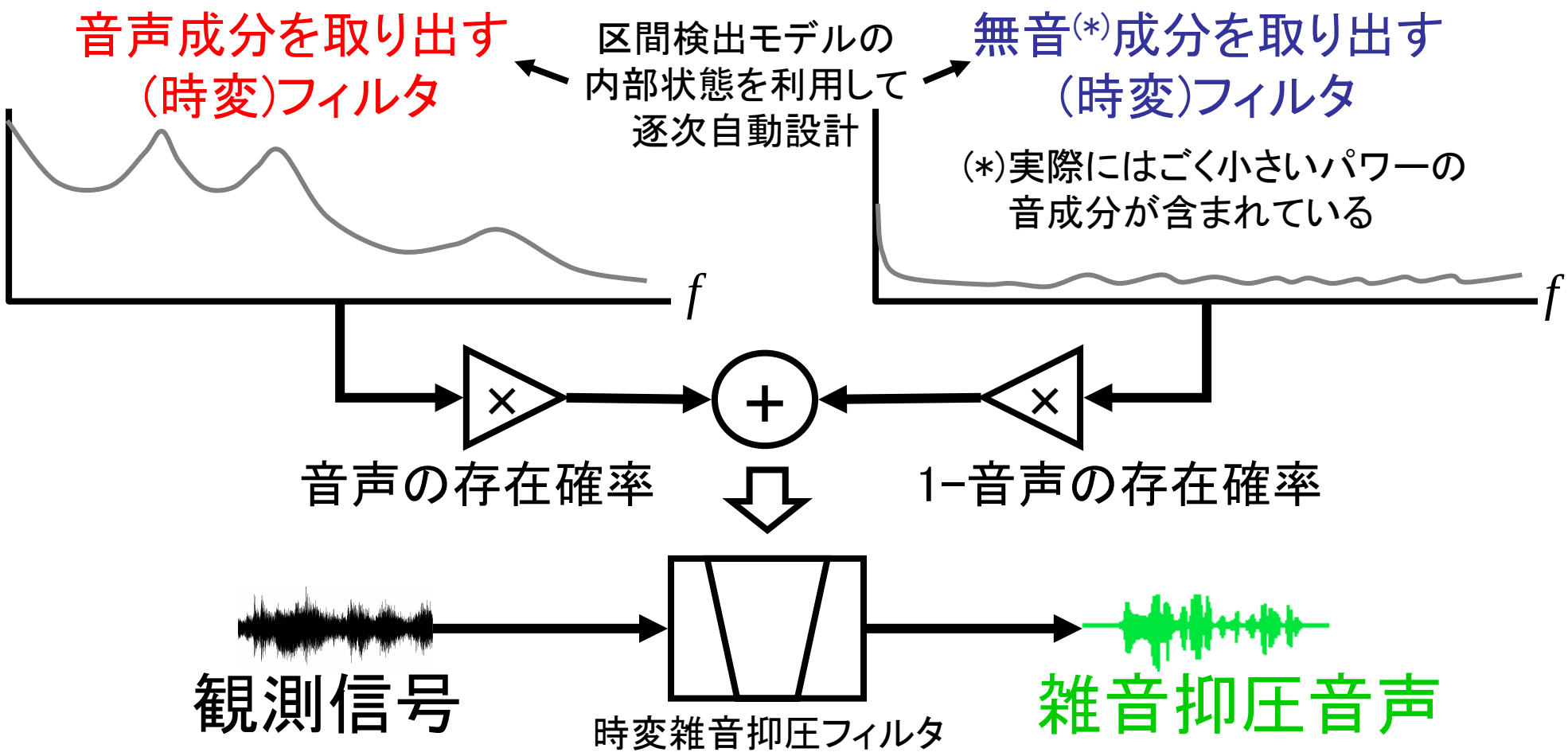
音声/非音声判定と同時に

各状態での**音声の存在確率**を計算

↳ **雑音抑圧に利用**

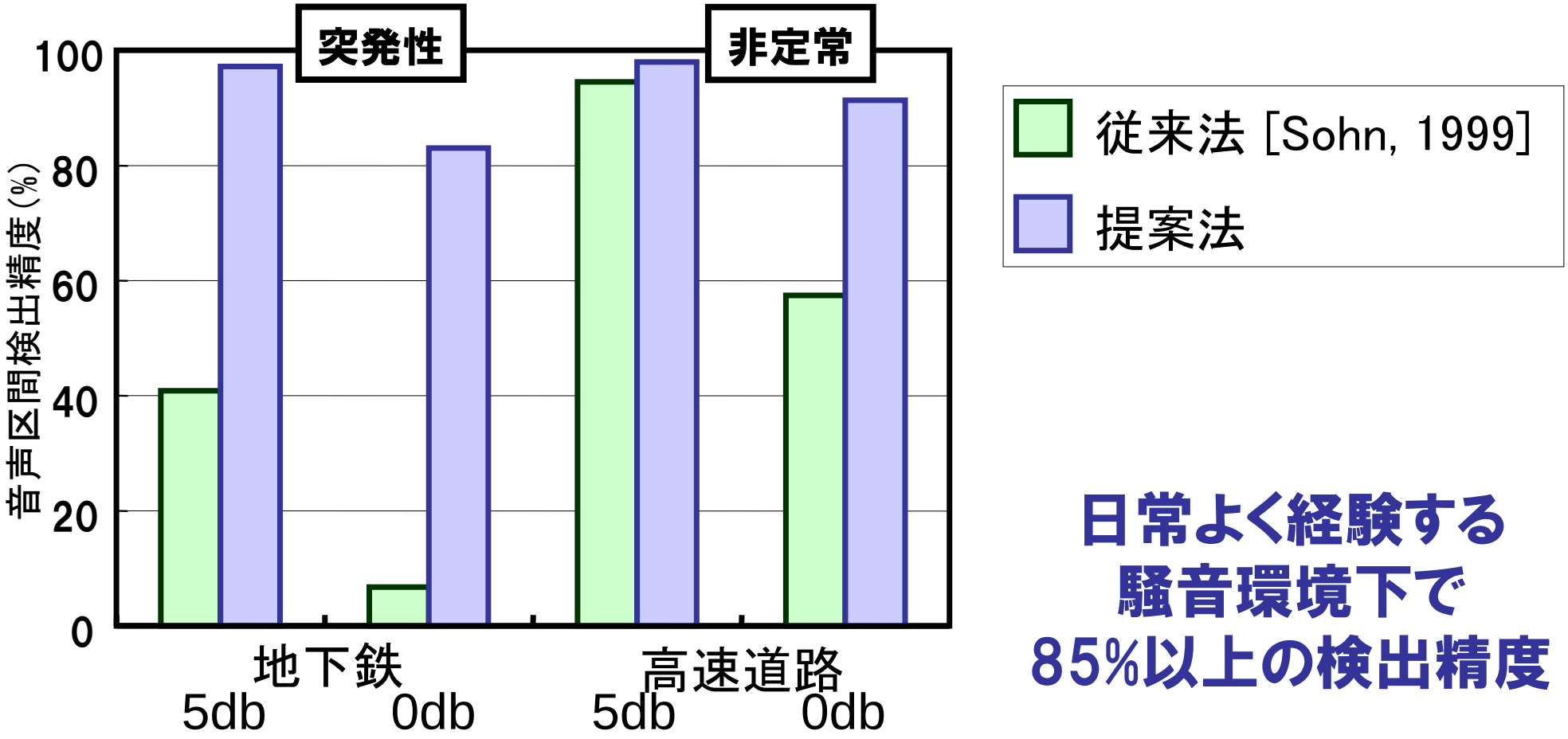
雑音抑圧技術の概要

各時刻の信号が音声である可能性の情報を利用した適応的な雑音抑圧



音声区間検出の効果

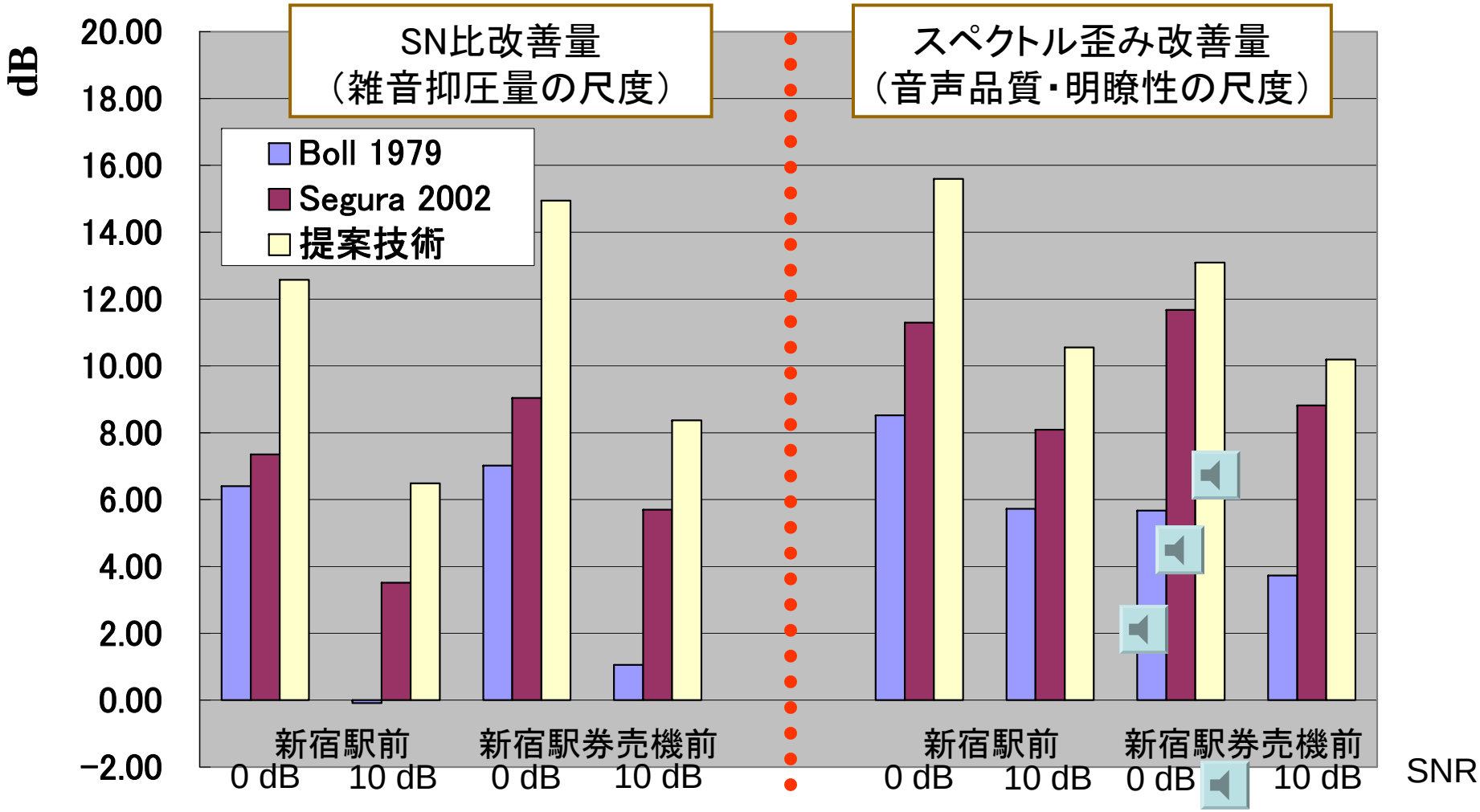
従来技術と比べ、大幅に高い検出精度



日常よく経験する
騒音環境下で
85%以上の検出精度

雑音抑圧の効果

従来技術と比べ、雑音抑圧、音声品質ともに改善量大



音声区間検出・雑音抑圧 –今後の課題・発展–

音声認識のフロントエンドとして、・・・

認識モデルとの密な連携・統合

汎用的な収音技術として、・・・

変化の速い雑音や突発性雑音の抑圧 

「雑音」の定義は凡そ主観的（信号音、情報音⇔雑音）

“目的の音を遮る、聞きたくない音のことを雑音と呼ぶ”

i.e. 人の声も、時と場合によっては雑音（オカンの小言）

⇒ 音信号中の聞きたい音声の時空間的位置を特定し聞きたくない音を取り除く技術へ

いつ、どこで、話したか？

（+ 誰が、何を話したか？（＝ 話者同定 + 音声認識））

「多様な環境下での音信号の音響・音声言語的解釈」技術へ

まとめ

コミュニケーション科学基礎研究所(CS研)における
NTTの音声処理基礎研究の一端を紹介

重み付き有限状態トランスデューサに基づく
超高速音声認識エンジン

音声区間検出＋雑音抑圧技術DIVIDE

「多様な環境下での音信号の音響・音声言語的解釈」技術へ