Vol.2013-SLP-98 No.3
2013/10/25

耐雑音音声認識エンジン VoiceDo の応用

服部　浩明[†1]　　辻川　剛範[†2]

VoiceDo は高騒音環境下でも高い精度で音声認識が可能な音声認識エンジンである．VoiceDo は製造現場や物流，医療，運輸などさまざまな騒音環境下でのデータ入力，コマンド入力に用いられている．ここでは VoiceDo で用いられている認識技術について解説するとともに，その適用事例を紹介する．また，あらたに適用領域を広げる試みとして，Bluetooth を用いた無線ヘッドセットと Android タブレットへの適用について紹介する．

Applications of VoiceDo: Speech Recognition Engine for Noisy Environments

HIROAKI　HATTORI[†1]　MASANORI　TSUJIKAWA[†2]

VoiceDo is a speech recognition engine with high recognition accuracy even in high noise environments. It has been adopted in various noise environments such as manufacturing premise, logistics, medical application, transportation, etc. Here, we describe about technologies used in the engine and show some application examples. In addition, we also describe a wireless headset using Bluetooth technologies and Android version of VoiceDo as a trial to extend the applicable area of speech recognition.

1. はじめに

音声認識技術の向上と情報機器の演算能力の向上，メモリの大容量化，高速ネットワークの普及に伴い音声認識の実用化がすすんでいる．カーナビゲーションシステムの操作やスマートフォンでの情報検索など日常的に広く音声認識が用いられるようになってきた[1][2].

NEC での音声認識の研究は 1960 年の音声タイプライターの開発を始めとし，1978 年には世界初の連続音声認識装置 DP-100 を製品化した．図 1 に DP-100 の写真を示す．この装置は荷物の仕分けなどで手がふさがっているときに音声でデータ入力ができるシステムであり，それ以来，NEC はエンタープライズ向けの音声認識装置を世に送り出し，さまざまな業種，業務へ音声認識を適用してきた．図 2 に音声認識の適用領域，事例を示す.

図 1 連続音声認識装置　DP-100
Figure 1 Speech Recognition System DP-100

音声認識を用いることで，作業しながらデータ入力することによる業務の効率化や，目を離さずに機器を操作することによる安全性向上が図れるが，これらの事例へ音声認識を適用する場合に最も問題となるのが環境雑音である．業務向けの音声認識システムはデータ入力や機器制御などを目的としており，高い認識性能が求められるが，製造工場や整備，保守の現場には，機器の運転音や打撃音などさまざまな雑音が存在し，音声認識性能を低下させる要因となっている．本稿ではそのような環境雑音下でも高い認識性能を持つ音声認識システム VoiceDo[3]で用いられている技術について説明し，その適用事例を紹介する．また，適用領域を広げる試みとして Bluetooth を用いた無線ヘッドセットと Android タブレットへの適用について紹介する．

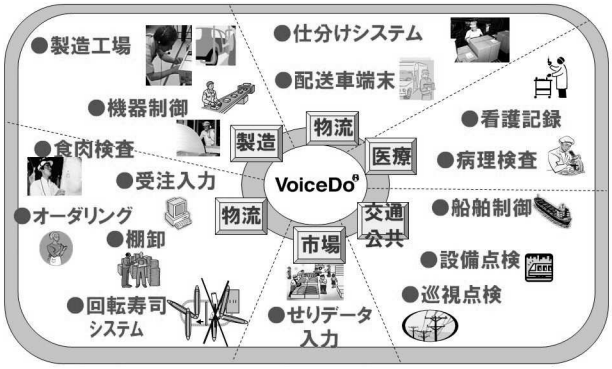

図 2 音声認識の適用領域，事例
Figure 2 Examples of speech recogniton application

†1 株式会社 NEC 情報システムズ
NEC Infomatec Systems, Ltd
†2 日本電気株式会社
NEC Corporation

2. システム構成

図 3 に VoiceDo の構成図を示す．VoiceDo には用途に合わせて PDA 版，PC 版が存在するが，エンジンの構成は同じである．以下それぞれの構成要素で用いられている技術について解説する．

2.1 音声検出，分析部

VoiceDo では音声マイクの他に雑音マイクを設置し，周囲雑音を推定し，音声マイクに回り込んだ雑音の抑圧を行う．この際，各マイクロホンで定常的な雑音を除去した後，雑音マイクの非定常な雑音を除去する２段階の処理を行っている[4]．また音声検出には，周波数方向にサブバンド化を行い，サブバンドごとに信号対雑音比 SNR を求め，最大の SNR を用いて検出を行うサブバンド音声検出[5]を行っている．これにより周囲雑音により一部の帯域がマスクされてしまうような高雑音環境下でも安定した音声検出を実現している．

2.2 音響モデル

音響モデルとして triphone の混合ガウス分布 HMM を使用している．一般に音響モデルはサイズが大きいほど認識精度が高まるが，その反面処理量が増加してしまう．そのため，PDA 等のリソースが限られた機器ではサイズを大きくせずに認識精度を確保することが求められる．そこで MDL 基準を用いた混合ガウス分布の削減[6]を行っている．この手法は，十分な数のガウス分布を学習した後，モデルの記述量を最少化するよう分布を削減していく方法であり，変動の少ない音素に対しては少数の分布を割り当て，変動の大きい音素に対しては多数の分布を割り当てる．削減の度合いは記述量のペナルティ係数を変化させることで制御でき，PDA 版， PC 版ではそれぞれに最適なサイズの音響モデルを作成している．

2.3 距離計算部

MDL 基準を用いた混合ガウス分布削減により距離計算対象となるガウス数を削減した後，さらに距離計算量を削減するために，木構造分布を利用した効率的な距離計算[7]を行っている．この方法では削減された混合ガウス分布を各分布間の類似度に基づいて木構造化しておき，認識時に親ノードから子ノードへ類似度の高い分布についてのみ距離計算を行い，類似度の低い分布については親ノードの距離で近似する．この方法により計算量を 1/10 以下に削減している．

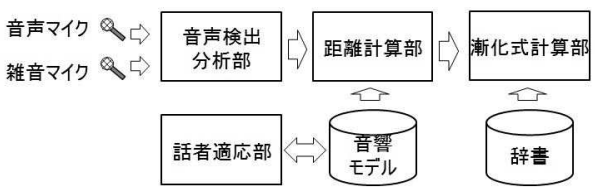

図 3 システム構成図
Figure 3 System configuration

2.4 辞書

辞書には各単語を構成する triphone の情報と，単語間の接続関係をネットワーク文法で記述した情報が格納されている．ここでもメモリ量を削減するために，各単語を先頭から同じ音素をマージした木構造型辞書[8]を採用している．辞書には認識対象をまとめた「ルール」を複数設定することが可能であり，認識時に有効とするルールを指定することで認識対象を簡単に変更することが可能となっている．これにより作業者名や日付等の入力項目ごとにルールを切り替えることで，認識精度を確保している．

2.5 漸化式計算部

データ入力や機器制御などの業務向け用途では，発声後に認識結果が返ってくるまでの応答時間が短いことが要求される．そこで，仮説探索にはワンパスのフレーム同期ビームサーチを用いている[8]．これにより，実時間での認識処理を可能とし，発話終了後 0.3 秒で認識結果を返す応答速度を確保している．

2.6 話者適応部

音響モデルは多数話者が発声した音声データベースから作成されており不特定話者での音声認識を可能としているが，業務向けにより高い精度を得るために話者適応機能を提供している．とくに周囲雑音の大きなところでは，話者の声質を学習するだけでなく周囲雑音の学習の効果もあるため，作業環境での話者適応を推奨している．

話者適応手法としては自律的話者適応方式[9]を用いている．この方法では木構造で表現されている音響モデルを適応する際に，適応用音声に含まれるデータ量に応じて適応する木構造中のノードを自動的に選択することで，データ量が多い場合も少ない場合も適切に話者適応が行われる．

表 1 に話者適応単語数を変えた場合の認識率を示す．話者は男女 10 名，語彙数は 1500 単語，雑音レベルは 80～85dB(A)である．適応単語数が 50 単語と少ない場合でも良好な認識性能が得られていることが分かる．

表 1 話者適応単語数と認識率
Table 1 Number of adaptation words and recognition rate

不特定話者	50 単語	100 単語	250 単語
90.8	96.4	97.1	97.9

3. 適用事例

ここでは VoiceDo の実際の適用事例についていくつか紹介する．

3.1 食肉検査場

食肉検査場では食の安全を確保するためにさまざまな検査が行われる．これまでは衛生上の観点から検査者とは別の記録者が検査結果を紙に記録していた．そこで検査作業

の効率化のため検査者が検査をしながら音声認識により検査結果をデータ入力できる検査システムを導入した．これにより，検査者一人での検査作業が可能になっただけでなく，ペーパーレス化による衛生面の改善，作業状況のリアルタイムでの把握が可能となった．

3.2 製品検査

各種工業製品の製造現場では出荷前の検査が行われる．近年，工業製品の多様化に伴い，多品種少量生産が求められるようになり，検査項目が多岐にわたるようになった．そのため，これまでの紙のチェックシートでは十分対応できなくなってきた．そこで音声合成により検査項目を指示し，検査を行いながら音声認識により検査結果を入力するシステム主導型の検査システムを導入した．これにより検査漏れがなくなっただけでなく，ハンズフリーでの作業が可能となり作業効率がアップした．

3.3 受注データ入力

大手小売業から卸売業への発注においては，流通 BMS (Business Message Standards)に代表される EDI（Electronic Data Interchange）の導入が進んでいるが，中小の小売店から卸売業への発注は，ＦＡＸや電話による方式がいまだに存在する．そのため受注する卸売業においては，ＦＡＸや電話で受けた手書き注文情報をデータ化するための入力業務に膨大な工数をかけている．この際，注文情報には卸売業の商品コードの記載はなく，類似する商品が膨大なこともあり，商品識別に多くの工数が必要となり，入力ミスも生じ易かった．そこで，音声認識を活用し手書き注文の商品名を読み上げることで該当する商品の識別を容易にする受注データ入力システムを導入した．これにより入力効率がほぼ倍となっただけでなく，入力ミスも約 7 件/日から約 0.03 件/日へと激減した．

4. 新しい適用領域開拓

ここでは新しい適用領域を開拓するために取り組んでいる新しい技術について紹介する．

4.1 無線ヘッドセット

これまでは PDA 版，PC 版ともにヘッドセットを有線で接続していた．そのため，作業中に線を製品等に引っかけて製品へ傷をつけたり，マイクロホンが断線したり，あるいはコネクタ部が損傷したり，といった不具合が生じることがあった．そのため，無線型のヘッドセットが欲しいという顧客の要望があり，昨年 3 月に無線型ヘッドセットを製品化[10]した．図 4 に写真を示す．音声マイクの他，筐体上部に雑音マイクを設置している．

図 4 Bluetooth 無線ヘッドセット
Figure 4 Bluetooth wireless headset

無線方式には Bluetooth を用いている. 携帯電話用の HSP（Headset Profile）を用いた無線ヘッドセットはパケットロスへの耐性が十分でなく，音声中にパケットロスによる無音が生じることがあり，音声認識に不向きである．そこでここではシリアル通信を行うための SPP（Serialport Profile）上に独自のプロトコルを構築し, パケットロスに対応した．SPP を介して，マイク２チャンネルの送信，スピーカ１チャンネルの受信と，電池残量の確認，音量制御等の機器制御を可能としている．サンプリング周波数は PDA 用の 11kHz，PC 用の 22kHz の二通りとし，通信量を削減するために波形の圧縮，伸長を行っている．

通信距離は 15m（class 1）あり，稼働時間は単三ニッケル水素充電池１本で約８時間と，製造現場等での利用に耐えるようになっている．

4.2 Android 版 VoiceDo

近年，iPad や Android タブレットなどタッチパネルの機能を持つ端末が安価で提供されるようになってきた．これまでのタッチパネル型 PC は高額であったため，VoiceDo と組み合わせて提案するとコストが高くなってしまう問題があったが，これらを利用することで音声認識ソリューションを安価に広い顧客へ提供できる可能性がある. そこで, Android タブレットへの VoiceDo 搭載の検討を行っている．

Android 端末に VoiceDo を搭載する場合，外付けのステレオマイク入力に対応している機種が少ないという問題がある．一方，キーボードやマウスを接続するための Bluetooth の SPP は Android 端末には広く搭載されており，前節の無線ヘッドセットを用いることで，さまざまな端末に対応することができる可能性がある．ここでは無線ヘッドセットと NEC の Lifetouch L[11]の組み合わせで試作を行った．Lifetouch L の仕様を表 2 に示す．CPU は 1.5GHz クロックのデュアルコア，メモリは 1 GByte と一昔前のデスクトップ並みの性能を持っている．

Android のアプリケーションは Java で記述されるが, Java

表 2 Lifetouch L 仕様
Table 2 Specification of Lifetouch L

プロセッサ	OMAP4460 1.5GHz (デュアルコア)
メモリ	LPDDR2 1 GByte
ストレージ	16～64GByte
OS	Android 4.0
ディスプレイ	10.1 型 1280×800 ドット
通信機能	802.11a/b/g/n 準拠，Bluetooth 2.1+EDR

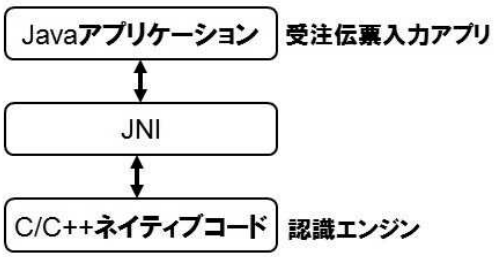

図 5 Android 版 VoiceDo の構成図
Figure 5　Configuration for Android

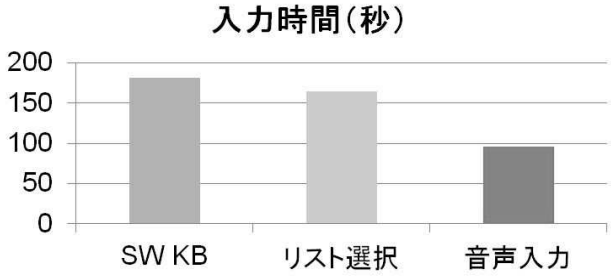

図 6　各種入力手段による伝票入力時間の比較
Figure 6　Comparison of input methods

の仮想マシン上で認識エンジンを動かした場合，十分な処理速度が出ない可能性があるため，認識エンジンそのものは C++のネイティブコードで実装し，JNI（Java Native Interface）を介して，Java アプリケーションから呼び出すこととした．図 5 にアンドロイド版 VoiceDo のシステム構成図を示す．

評価用アプリケーションとして，受注伝票入力アプリを作成した．入力するのは商品名（50 種）と商品コード（4 桁数字），数量（1～999），産地（47 都道府県）である．このアプリケーションは入力手段の比較ができるよう，ソフトウェアキーボード入力，プルダウンリスト入力，音声入力の３種類の入力が可能となっている．評価は紙に書かれた１０枚の注文票をそれぞれの入力方法で入れた場合の入力時間を比較した．その結果を図 6 に示す．音声認識は他の入力手段に比べ，半分程度の時間で入力可能であり，音声認識の有効性が確認できた．

プルダウンリストの方がソフトウェアキーボードより早いと予想していたが，それほど大きな差は出なかった．これは Android の IME（Input Method Editor）の単語予測の精度が良く，最後まで入力せずに正解候補が得られたことと，プルダウンリストでスクロールさせて項目を選ぶ場合，スクロールした画面内に入力したい項目があることを確認するために時間がかかったと考えられる．

上記比較では音声認識誤りが生じていないが，実際の利用場面では音声認識誤りが生じるため，誤り訂正の時間を含めた比較が必要である．一般に音声認識誤りを音声認識のみで修正するのは修正コマンド自身を誤認識する可能性があるため難しい．そこで，ここでは音声入力を行い，誤った場合にはタッチパネルで修正する場合を考える．

ソフトウェアキーボードやプルダウンリストの入力時間を 1 とし，音声認識の入力時間をその半分の 0.5 とする．簡単のために個々の項目の入力時間は等しいものとし，音声認識精度は表 1 から 97%とすると，修正にかかる時間は 0.03 となる．したがって，音声認識誤りの訂正を含めた入力時間は 0.53 となり，音声認識とソフトウェアキーボード/プルダウンメニューを組み合わせることで快適な入力手段を提供できると言える．

5. おわりに

耐雑音音声認識 VoiceDo で用いられている技術について説明し，その適用事例を紹介した．VoiceDo は業務向けの音声認識エンジンであるが，近年 B2C 領域でのトライアル[12]も行われており，今回紹介した無線ヘッドセット，タブレットへの取り組みを含め幅広く適用領域を探して行きたい．

参考文献

1) NTT ドコモ：しゃべってコンシェル http://www.nttdocomo.co.jp/service/information/shabette_concier/ .
2) Apple: Siri, http://www.apple.com/jp/iphone/features/siri.html.
3) NEC：VoiceDo，http://www.nec.co.jp/voicedo/.
4) 高木，吉田，渡辺, "2 段スペクトルサブトラクションによる雑音化音声認識," 音講論, pp.59-60, 2-5-3, 1991 年 3 月.
5) 江森，辻川，大西，越仲，谷，北出，佐藤，“法廷音声認識システムの開発 －複数マイクロフォンを用いた音声検出－”，音講論, pp.41-42, 1-6-16，2010 年 3 月.
6) K. Shinoda, T. Watanabe. MDL-based context-dependent subword modeling for speech recognition, Journal of Acoustic Society of Japan (E), Vol. 21, No. 2, pp. 79-86, 2000.
7) 渡辺，篠田，高木，山田，服部，磯, "木構造確率分布を用いた音声認識," 音講論, pp.13-14, 1-8-7, 1993 年 10 月.
8) 友枝，石川，大川，江森，磯，“木構造辞書とネットワーク文法を用いたコンパクト大語い連続音声認識エンジン,”音講論, pp9-10, 2001 年 3 月.
9) 篠田，渡辺，“音声認識における自律的なモデル複雑度制御を用いた話者適応化,”信学論 D, Vol.J79-D2, No.12, pp2054-2061.
10) NEC: newsrelease, http://www.nec.co.jp/press/ja/1203/1901.html.
11) NEC: Lifetouch L, http://121ware.com/lt/.
12) 花沢，辻川，“キャラクターとの会話体験を提供する音声応答の試験サービス,”情報処理学会研究報告，Vol.2012-SLP-93, No.8，2012 年 10 月.