スマートフォン向け日中英音声翻訳システム

Japanese/Chinese/English Speech-to-Speech Translation System for Smartphones

井阪　岳彦　　知野　哲朗　　永江　尚義
■ ISAKA Takehiko　■ CHINO Tetsuro　■ NAGAE Hisayoshi

外国語によるコミュニケーションの必要性がますます高まっており，最近は英語だけでなく中国語の音声入力による自動翻訳で自由に意思疎通したいユーザーが急増している。

東芝は，スマートフォン(注1)向けに音声入力による自動翻訳システムを開発した。このシステムでは，雑踏の中でも話しことばを正確に音声認識し，自然な表現に翻訳して聞き取りやすい合成音を生成する。スマートフォン単体で日本語，中国語，及び英語それぞれの間の音声自動翻訳を実現しており，一般的な文章であれば３秒程度で音声翻訳でき，母国語が通じない海外渡航先でもコミュニケーションの世界を広げていくことができる。

Toshiba has developed an automatic speech-to-speech translation system for smartphones that realizes all-directional interpretation between Japanese, Chinese, and English as a standalone system. This system can recognize speech accurately even in noisy environments, translate it into natural expressions, and synthesize clear speech. The typical time required for both speech recognition and machine translation is within about 3 seconds. This system will enable people in various countries to broaden their communication horizons.

1　まえがき

最近，グローバル化の流れはますます加速しており，外国語による意思疎通の必要性が高まっている。英語圏はもちろんのこと，経済発展の著しい中国語でのコミュニケーションも必要になってきた。

人類にとって長年の夢であった音声入力による自動翻訳は，既に携帯電話などで商用サービスが開始されている。しかし，サーバで実現する音声翻訳システムでは，ネットワークへの接続コストや，応答速度，確実性などの面で課題があった。

東芝は，このような課題解決のニーズに応えて，日本語，中国語，及び英語それぞれの間で音声翻訳ができるシステムを開発し，当社のスマートフォンで動作することを確認した。ここでは，このシステムの概要と特長について述べる。

2　日中英音声翻訳システムの概要

実用的な音声翻訳システムでは，ユーザーが自分の意思を自由に伝えられることが重要である。音声翻訳システムの画面例を**図１**に示し，以下に概要を述べる。

まずユーザーのボタン操作をトリガ（きっかけ）として，スマートフォンの内蔵マイクから音声を取り込む。この音声に対して，音声認識を実施し，その結果のテキストを画面上に表示する。次に認識結果を翻訳して，その結果のテキストを画面上に表示する。最終的に，翻訳結果から合成音を生成して，内蔵スピーカで再生する。

図１．音声翻訳システムの画面 ― 音声入力を相手言語に自動的に音声翻訳する。

Example of speech-to-speech translation system display

画面中央のボタンを押すことで，翻訳方向が瞬時に切り替わり，会話の翻訳を円滑に支援する。音声入力中は入力音量がメータで表示されるので，入力状況を把握できる。事前に声を登録しなくても，簡単に音声翻訳ができるように配慮して

（注１）　携帯電話とパソコンや携帯情報端末（PDA）を融合させた携帯端末機器。

いる。ネットワークに接続する必要がないため，通信できないエリアでも確実に音声翻訳でき，通信費を気にせず，海外でもコミュニケーションの不安や不便を軽減できる。

3　システムの構成

開発した音声翻訳システムの構成を**図2**に示す。このシステムは，基本的に以下の三つのモジュールから構成されている。

(1)　周辺雑音に強く，多言語に展開できる“音声認識”

(2)　話しことばが持つ曖昧（あいまい）性を理解できる“機械翻訳”

(3)　自然で肉声感のある音声を生成する“音声合成”

これらのモジュールを小型・高速化したことで，話しことばのような曖昧性のある会話文を，雑踏の中でも正確に音声認識し，自然な表現に翻訳して，聞き取りやすい合成音を生成する。以下，各部とこれらを操作するためのGUI（Graphical User Interface）部について述べる。

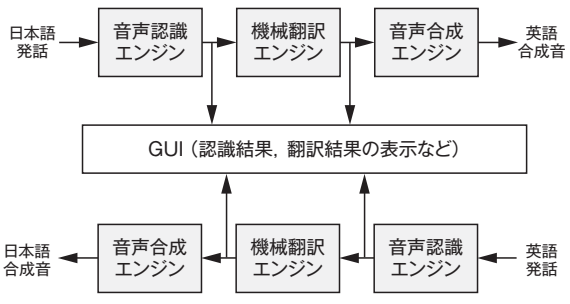

図2．音声翻訳システムの構成 ― 2言語の間で双方向に，音声認識で発話をテキストへ変換し，機械翻訳で相手言語へ翻訳した後に，音声合成で合成音を生成する処理を行う。

Configuration of speech-to-speech translation system

3.1　音声認識

ユーザーが発話する様々な文章を効率よくカバーして音声認識するため，ここでは，3万語程度の単語を組み合わせて，文章を認識できる連続単語音声認識技術を開発した。単語認識から大語彙（ごい）の連続単語認識まで，スケーラブル（汎用性がある）かつコンパクトな音声認識を実現している。多言語を想定した設計となっており，ここで開発した日本語，中国語，英語をはじめ，世界主要国の言語に共通して展開できるように配慮している。更に，車載向け音声認識システムの開発で培った雑音対策技術を採用しており，音声翻訳システムが実際に使用される場面でも高い認識精度を実現している。

音声認識エンジンは，**図3**に示すように，音響分析と音声認識デコーダ（音響特徴量からテキストへの変換）の二つのモジュールで構成されている。音響分析部では，入力音声信号の中から発話されている区間を検出し，音声認識に必要な情報（音響特徴量）を抽出する。音声認識デコーダ部では，音響特徴量を認識単語の組合せとして文章に変換する。

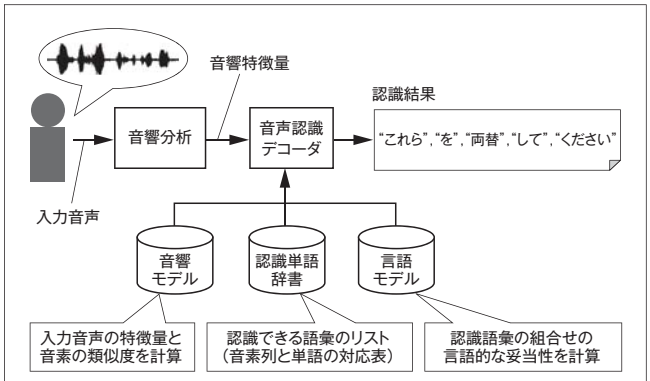

図3．音声認識エンジンの構成 ― “入力音声に対する音響的な類似性”，“音素を並べて生成した単語列の言語的な妥当性”の二つの結果に基づいて最適な認識結果を生成する。

Configuration of speech recognition system

音響分析部には，声であるか否か（認識すべき対象であるか）の識別精度を高める技術と，周辺雑音を抑圧して認識精度を高める技術を採用している。

音声認識デコーダ部は，音響モデル，認識単語辞書，及び言語モデルの三つのデータベースを併用して，発話内容を高精度でテキストに変換する。ここで，音響モデルとは，音素の音響的な特徴量をデータベース化したものである。事前に声を登録しなくても不特定の話者で音声認識でき，耐雑音性が高い独自の音響モデルを採用している。音響的な特徴量をデータベース化する際に，話者によって変動しにくい特徴量に注目することで，話者ごとの認識精度のばらつきを抑えている。また，雑音免疫学習法[1]を採用することで，各音素の雑音下での峻別（しゅんべつ）力を高めている。

一般に音声認識精度を高めるためには，膨大な数の単語を認識単語辞書としてシステムに登録しておく必要があり，システムで大きなメモリ量を消費していた。ここでは，音声翻訳システムが旅行で使われることを想定し，旅行中に使われやすい約3万の語彙に絞ることで，未知語の発生頻度を抑えながら，実用的でコンパクトな認識単語辞書を開発した。

更に，言い回し情報（言語モデル）を用いて，認識された単語の組合せに対する言い回しの妥当性を計算することで，音の類似性と文章としての妥当性の両方の観点からもっとも適切な文章を認識結果として出力する。このデータベースは，事前に用意した例文集からN-gramと呼ばれる単語間の出現頻度パターンを集め，統計的な確率情報をまとめたものである。ここでは，例文集として旅行会話集を重点的に集めて，旅先の会話で使われる文章の傾向をデータベース化することで，高い認識精度を実現している。

これらの技術を結集することで，幅広い人に対して事前に声

を登録しなくてもすぐに使用でき，雑音環境でも高い認識性能を維持した実用的な旅行会話の音声認識エンジンを実現している。

3.2 機械翻訳

入力言語ごとに形態素，構文，係受け，意味などの解析処理をした後，出力言語に言語構造を変換し，最適な単語を選択して，最終的に出力言語のテキストを出力する。日英辞書などの単純な単語の置換えや，旅行会話集などの文例集とは異なり，任意の文章を翻訳できる。

ここでは，“スマートフォンの処理能力”と“必要となる精度”のバランスに基づき，“規則翻訳方式”と“統計翻訳方式”を翻訳方向ごとに選択し，精度の高い翻訳エンジンを実現した。これらの翻訳エンジンのうち，口語処理に特化した統語森駆動方式という当社独自の規則翻訳エンジン[2]の概要を述べる。

図4に示す統語森構造を処理の基盤とし，音声翻訳の様々な過程で生じる曖昧性を一括して表現し，同時処理ができるようになった。

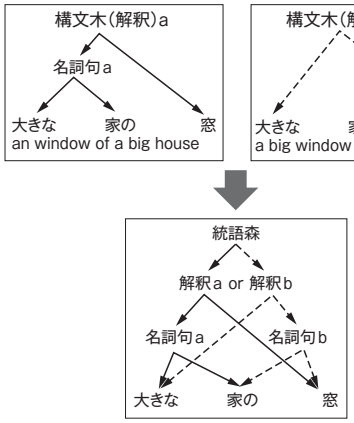

図4．統語森構造の概要 — 統語森構造は，解析の過程で生じる曖昧性に対応する複数の構造候補の共通部分を共有し，一つのデータとして圧縮して表現する知識表現方法である。

Outline of syntactic forest structure

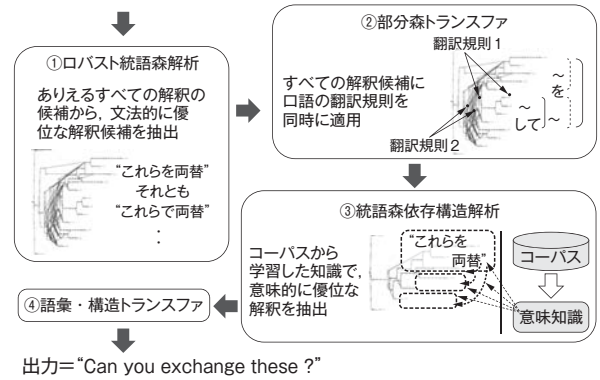

図5．統語森駆動規則翻訳処理の概要 — 話しことばに固有の言い回しを文法解析して，最適な規則の組合せを適用する。

Outline of translation processing based on syntactic forest-driven rules

統語森駆動・規則翻訳処理の概要を**図5**に示す。まず，音声認識エンジンから得られる認識結果（入力）に対して，ロバスト統語森解析処理（①）が施され，続いて部分森トランスファ処理（②）が施される。次に統語森依存構造解析（③）が施される。最後に，語彙・構造トランスファ処理（④）によって，訳文（出力）が生成される。

前述の統語森構造をベースとした①から③の処理によって，話しことばに起因する曖昧性を効率よく解決して，話しことばと同様に翻訳できるようにしている。

3.3 音声合成

音声合成は，任意の入力テキストを音声に変換する技術である。大量の音声データを用い，統計的な手法に基づいて音声を合成する“コーパスベース音声合成”を採用することで，コンパクトなメモリサイズで安定した高品質な音声合成を実現している。この技術は，**図6**に示すように，テキスト解析，韻律生成，及び音声信号生成の三つの処理から構成される。

テキスト解析部では，言語辞書を参照して，入力されたテキストを解析する。日本語を例として説明すると，漢字の読みやアクセントの位置，文節（アクセント句）の区切りなど言語情報を出力する。

韻律生成部では，言語情報に基づいて，声の高さ（基本周波数）の時間変化パターンと各音韻の長さなどの音韻・韻律情報を出力する。ここでは，音声データ（音声コーパス）から抽出された自然音声の基本周波数パターンを教師データとして制御規則を自動的に学習する。この手法では，アクセント句単位の典型的な基本周波数パターンを表す代表パターンに基づくモデルを導入している。このモデルから出力される基本周波数と実音声の基本周波数の誤差を評価尺度として，誤差を最小化するように学習する[3]。これにより，話者の特徴をとらえた基本周波数パターンの生成を実現している。

音声信号生成部（合成器）では，音韻の系列に従って音声データ（音声素片）を選択し，韻律情報に従って変形して接続することで，合成音声を生成する。ここでは，独自の合成器で

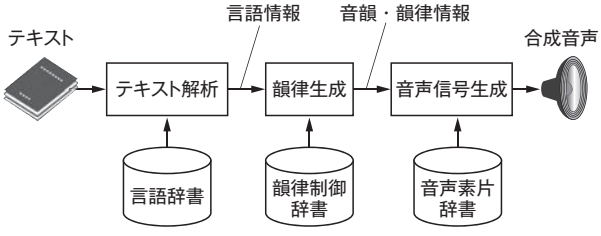

図6．音声合成エンジンの構成 — 任意の入力テキストから，テキスト解析，韻律生成及び音声信号生成の3段階の処理を経て，合成音声が生成される。

Configuration of text-to-speech conversion system

ある複数素片選択融合方式を採用した。この方式では，音声データから単位当たり複数個の音声素片を選択し（複数素片選択），それらを融合して合成単位の音声素片を生成する（素片融合）。適切な音声素片がコーパスに存在しない場合でも，目標近傍の複数の音声素片を融合することで，目標に近い特徴を持った音声素片を作り出して，均質で肉声感のある安定した音声を生成している。

複数素片選択融合方式では，従来の手法に比べて，音声素片融合の計算量が増加する。そこで，あらかじめ大量のテキストを合成し，選択された音声素片の組合せごとに出現頻度を記録して，高頻度の組合せだけ事前に融合音声素片を作成しておく。音声コーパスの代わりに融合音声素片を利用することで，音質をほぼ維持したまま，大幅な計算量の削減を実現している。

3.4 GUI制御

ユーザーのボタン操作に従って，音声入力から，音声認識，機械翻訳，音声合成，合成音の再生といった一連の音声翻訳処理を実行する。認識結果のテキストを画面に表示して，システムに正しく音声が入力されたことを発話者にフィードバックする。翻訳結果は，合成音とテキストで提示できる。例えば，航空機内でほかの搭乗者が就寝中のときには，音量を絞ってテキストだけで提示するといった使い方ができる。状況に応じて出力方法を切り替えられるので，実用的である。

音声入力方法は，人によって個人差が出やすく，使いやすさ，音声認識精度に大きく影響する部分である。ボタンを押し離してから音声入力する人や，ボタンを押したままの状態で音声入力する人，ボタンを押しいったん離してからボタンの上に再び指を置いた状態で音声入力する人など，様々である。ここでは，音声入力方法を事前に設定しなくても，これらの音声入力にすべて対応できるようにシステムを設計した。

ことばの通じない相手がシステムを操作する場合も想定し，初めて操作しても使いやすいように，画面構成を設計した。ここでは，メンタルモデルを意識して，画面の上から下に音声翻訳の処理が流れるように配置した。ボタン部分の表示言語などを音声入力言語に合致させる，入力音量に連動して音量メータの表示を更新する，翻訳方向を切替える際にアニメ表示するなど，実用性とデザイン性が両立するように設計した。

4 スマートフォンでの音声翻訳技術の評価

当社製のスマートフォン上に，これらの音声翻訳技術を結集し，旅行会話についてサーバ型に匹敵する音声認識率と，翻訳品質，合成音質を実現した。簡単な文章であれば，音声を入力してから約3秒で音声認識と機械翻訳の処理を完了し，認識結果と翻訳結果のテキストを画面上に表示し，音声合成による読上げを開始できる。国内外のフィールド評価で，様々な周辺雑音のある中でも日本語，中国語，英語の相互で実用的に音声翻訳できることを確認した。

開発したシステムを，映像，情報，及び通信の国際展示会であるCEATEC JAPAN 2009に出展し，コミュニケーションの可能性を広げられるシステムとして，米国記者からメディアパネル イノベーション アワードのファイナリストを受賞した。また，MWC（Mobile World Congress）2010では，関係各方面から引合いがくるなど，世界中から高い評価を受けた。このシステムは，当社製スマートフォン“dynapocket$_{TM}$”（ドコモ“T-01B”及びau“IS02”）ユーザー向けに“ポケット通訳$_{TM}$”として一般公開された。

5 あとがき

当社は，スマートフォン向けに音声入力による実用的かつ経済的な自動翻訳システムを開発した。話しことばのような曖昧性のある会話文を，雑踏の中でも正確に音声認識し，自然な表現に翻訳して，聞き取りやすい合成音を生成できる。

今後，音声認識及び機械翻訳の精度改善と，音声合成の音質改善，これらの多言語化及び多様化によって，コミュニケーションできる言語のバリエーションと質を更に高めて，自動翻訳システムの顧客価値を増大していく。

文 献

(1) 竹林洋一，ほか．ワードスポッティングによる音声認識における雑音免疫学習．電子情報通信学会論文誌．J74-D-Ⅱ，2，1991，p.121－129.
(2) 知野哲朗，ほか．日中英3言語6方向音声翻訳システム．情報処理学会第185回自然言語処理研究報告NL-185．2008，46，2008，p.15－22.
(3) 籠嶋岳彦，ほか．代表的パターンコードブックを用いた基本周波数制御法．電子情報通信学会論文誌．J85-D-Ⅱ，6，2002，p.976－986.

井阪 岳彦 ISAKA Takehiko
ビジュアルプロダクツ社 コアテクノロジーセンター モバイル技術開発部主務。音声信号処理技術の開発に従事。日本音響学会会員。
Core Technology Center

知野 哲朗 CHINO Tetsuro
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。自然言語処理，ヒューマンインタフェースの研究・開発に従事。情報処理学会会員。
Knowledge Media Lab.

永江 尚義 NAGAE Hisayoshi
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。音声認識，言語モデルの研究・開発に従事。情報処理学会会員。
Knowledge Media Lab.