Technology Reports

CSR　音声認識　通話音声テキスト化

聴覚に障がいのある方の生活を支援する「みえる電話」のサービス検証と開発

ドコモ・テクノロジ株式会社　パケットNW事業部　三上 和愛† 篠﨑 卓也
サービスデザイン部　森田 潤介

ドコモでは，すべての人が使いやすい製品・サービスの提供というCSRの観点から聴覚に障がいのある方や聞こえづらい方向けに通話音声をリアルタイムにテキスト化してスマートフォンの画面に表示するサービスを提案した．システムおよびアプリ開発においては，聴覚に障がいがある多くのユーザの意見を聞き，簡易なプロトタイプをベースに，アプリ操作性や音声認識エンジンのチューニングなどの仮説検証を繰り返した．

1. まえがき

障がい者差別解消法（2016年4月1日施行）により，合理的配慮の下，障がい者向けサービスの機能拡充が求められる社会である日本において，高齢者を含め，電話の通話音声が聞き取りにくいという方は700万人以上いると言われている．

インターネット社会となり，Webサービスが普及してきたが，「電話問合せ」や「電話申込み」など，電話での連絡のみ可能な場面もまだまだ多く存在しており，この状況は聴覚に障がいがある方にとって生活の妨げとなる．実際に，聴覚に障がいのある方へアンケートを行った結果，「聴覚に障がいがあることで困ること」として最も多かったのは「電話が必要なシチュエーション」（58.1%）であった．

特に，クレジットカードの紛失や，水まわりの故障などのライフライン上のトラブルといった緊急時においては，音声通話でないと解決できず，非常に困惑するという実態が明らかになった．

既存のサービスとして，オペレータが伝達を仲介するものも存在するが，利用可能時間帯が限られており，かつコストもかかるため積極的には利用されていないのが現状である．

† 現在，ソリューションサービス事業部

一方，音声認識の技術成熟により，通話音声のリアルタイムテキスト化が実現できる見込みが立ってきたことから，ドコモでは聴覚に障がいのある方の生活を支えるサービスとして，通話音声をリアルタイムにテキスト表示する「みえる電話」のサービス検討を開始した．本サービスを検討する上で，通話の相手側は音声認識を意識した話し方ではないため，音声認識精度が低くなるという課題があり，サービス提供可能な音声認識精度という点で，聴覚に障がいのある方との通話コミュニケーションが成立するかを確認する必要があった．そこで聴覚に障がいのある方を対象としたAndroid™*1アプリによるプロトタイプ検証により，サービスコンセプトの確認／現状の認識精度でのユーザ評価の確認／必要とされるミニマム機能の抽出を行った．

その結果，認識精度が向上すればぜひ利用したいという声が多く，利用シーンとしては友人や家族同士の通話よりも，企業への問合せなど通話の相手が知人以外の電話の場合に多く利用するという事実も判明した．

それを踏まえ，利用シーンを考慮した機能面での拡充と認識精度の向上を図り，それらを適用したトライアルサービスとして提供した結果，サービス性や品質で高い評価を得ることができ，商用サービスとして提供するまでに至った．

本稿では「みえる電話」トライアルサービスにおける通話音声テキスト化の実現方式および音声認識の精度向上施策内容と，商用サービスを提供するために開発した専用アプリ・システムについて解説する．

2. トライアルサービス開発

2.1 概　要

通話音声テキスト化に対する需要数把握および利用者満足度の測定，音声認識精度の向上を目的にモニターユーザを募り，「みえる電話」のトライアルサービスを，2016年10月から提供開始した．

「みえる電話」トライアルサービスの概要を**図1**に示す．トライアルサービスの設計として以下の提供機能に関する要求条件を定義した［1］．

①リアルタイム性の要件

通話中，通話相手の音声を音声認識し，サービス利用者のスマートフォン上にテキストとしてリアルタイムに表示できること．また，通話開始可能となるタイミングをサービス利用者が把握できるよう，通話状態をスマートフォンの画面に表示すること．

②端末非依存の要件

サービス利用者の端末はスマートフォンであればOSに依存することなく幅広い機種で利用が可能であること．通話相手側の端末は，音声通話が可能な電話機であれば，（アプリなど無しに）利用可能であること．

③法的配慮の要件

通話音声を録音・テキスト化すること，および，録音音声をサービス性向上のために利用する場合があることをサービス利用者に説明し，サービス利用者から同意を得る機構を有すること．また，通話相手側へも通知し，プライバシーの配慮が行えること．

2.2 通話音声テキスト化サービス実現方式

前述の要求条件を満足する実現方式としてNWサービス方式を採用した．NWサービス方式は，音声通話路上に通話音声の録音が可能なメディア処理装置を構成し，通話相手の発話をメディア処理装置で録音後，録音音声を音声認識エンジン*2へ転送，音声認識結果であるテキストを音声認識エンジンか

＊1　Android™：スマートフォンやタブレット向けのオペレーティングシステム，ミドルウェア，主要なアプリケーションからなるソフトウェアプラットフォーム．米国Google, LLC.の商標または登録商標．

＊2　音声認識エンジン：音声データを入力し，発話内容をテキスト化する装置．

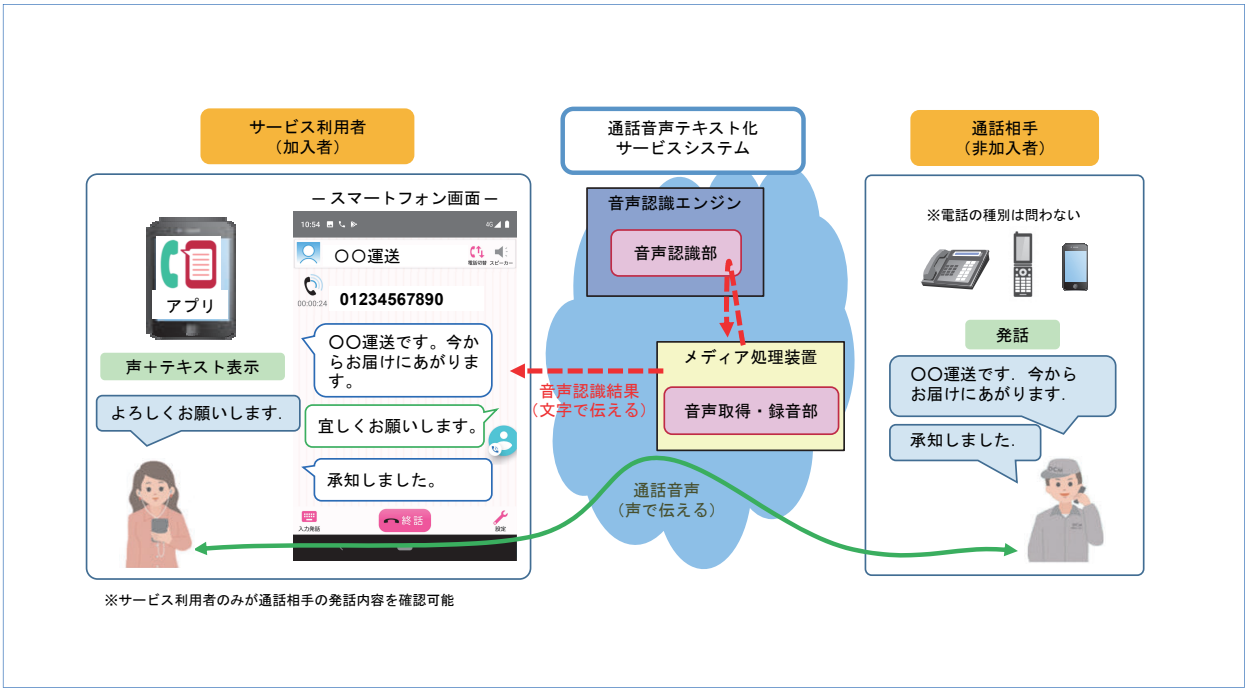


図1 「みえる電話」トライアルサービスの概要

ら受け取る方式である．トライアルシステムアーキテクチャを**図2**に示す．

①リアルタイム性の確保

メディア処理装置が通話中に無音を検知したタイミングで通話音声の録音と音声認識を行い，リアルタイムに音声認識結果を表示できるようにした．

なお，ガイダンスが終了して通話可能となるタイミングをサービス利用者が認識できるよう，スマートフォン画面に通話状態を表示することも可能とした．

②端末非依存性の確保

音声通話路上で通話音声を録音するため，端末での録音機能の実装を不要とし，音声通話機能と簡易なテキスト表示部のみを端末配備とするアーキテクチャを構成した．具体的には，音声通話機能は端末搭載の電話アプリを使用し，Android OS向けには専用アプリを開発し通話音声テキスト表示部を実装した．

また，その他のOS向けに，専用アプリを使用しない場合でも利用できるようにするため，サービス処理装置にOS非依存のWebアプリを構築し，標準ブラウザ画面上でのサービス機能の利用を可能とした．

これら実装により，機種やOSに依存せずにスマートフォンであれば利用可能とした．

なお，音声通話路上で通話音声を録音する構成としたことにより，通話相手の端末は音声通話が可能な電話機であれば（アプリなど無しに）利用を可能とした．

③法的配慮の実現

通信の秘密に関する同意取得については，通

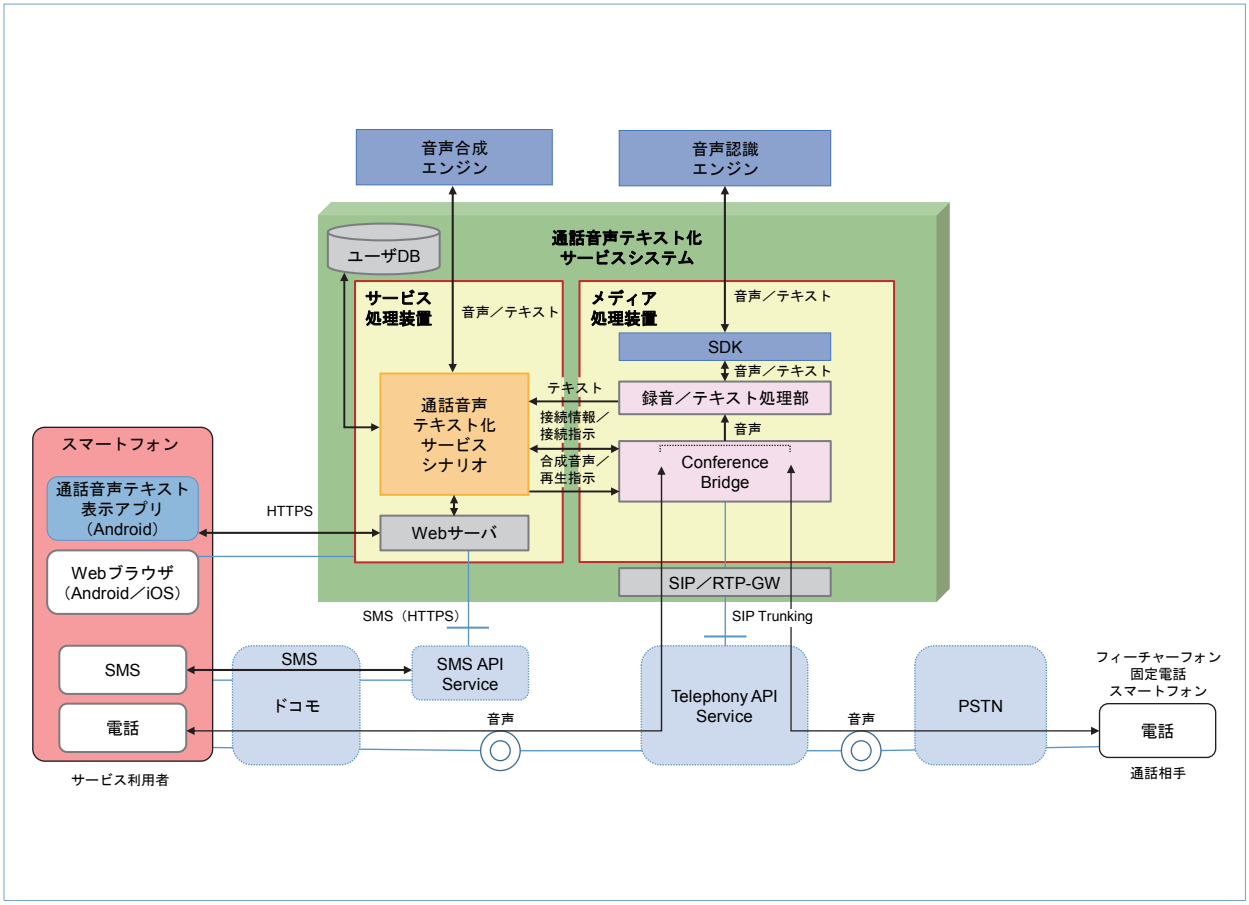


図2　トライアルシステムアーキテクチャ

話内容をテキスト化に利用する旨，サービス利用者の専用アプリ（またはWebアプリ）画面に表示し，「同意」ボタンを押下することで，個別明確な同意取得を行える構成とした．

また，通話相手側へのプライバシー配慮については，着信側の受話と同時に（またはサービス利用者による手動契機により），通話が録音されテキスト化に用いられる旨を通知する音声ガイダンスをメディア処理装置から通話相手へ送話する構成とした．

2.3　音声認識精度向上の取組み

⑴音声認識精度向上のためのガイダンス・チューニング

「みえる電話」は，テキストの正確性，つまり音声認識精度によってサービスの使い勝手が大きく変わる．どのような会話でも完璧に変換することは技術的に難しいため，現時点では認識精度が高まるような工夫をしつつ，利用シーンに合わせた現実的な範囲の目標を定義し改善に取り組んでいる．具体的には，それまで通話が難しかった難聴者をサポートすることを目的としているため，まずは「テキスト変

換結果から読み手が次の行動を判断できるレベル」をめざしている．また，音声認識精度を向上させる取組みとして2方向からアプローチしている．1つは利用シーンとガイダンスによって，認識しやすい音声になる確率を高めること，もう1つは実際の発話データを利用した音声認識エンジンのチューニングである．

「みえる電話」が主な利用シーンとしている「電話が必要なシチュエーション」では通話相手が公的機関やお店，企業の問合せ窓口であることが多い．このようなシーンでは認識しやすい明瞭な発話がされるため，音声認識精度が高くなる傾向がある．加えて，通話開始時に「音声認識を利用するのではっきりと発話する」ことを促すガイダンスを流すことで，明瞭な発話の意識付けを図っている．

音声認識エンジンのチューニングについては，トライアルサービスの提供と並行して定期的に音声ログを用いた音声認識精度向上を行ってきた．ユーザの同意を得て使用した音声を収集して分析し，店舗名などの頻出単語や使用されたシチュエーションに関連する単語を音声認識エンジンに辞書登録している．

さらに，発話文章を音声認識エンジンが学習することで利用シーンへの最適化が進み，その結果トライアルサービス提供開始当初と比較して，文字正答率で10%弱の音声認識精度の向上が見られた．本格商用サービスが始まり利用者が増えることで，より多くの音声サンプルが集まるため，より効果的な精度改善の手法を取ることができると考えている．

⑵連続認識方式の改善

通話音声の場合，リアルタイムに連続して音声認識する必要があるため，当初はサービス処理装置で音声通話の開始を契機として録音・音声認識を自動的に開始し，無音検出を契機として録音・音声認識を停止した後，続けて再開させる手順を構成した［2］．そこでは，無音検出から次の録音開始までの間，録音をしない時間が生じるものの，録音欠落時間は数十ミリ秒程度に収まり，通話音声の無音時間内になるだろうと想定していた．しかし，実際にシステムを構築し，試験を実施したところ，数百ミリ秒程度の録音欠損時間が発生し，それに起因して文頭切れ（発話の最初の文字が録音されない），音声認識結果が悪い（文頭切れにより音声認識されない）という2点の問題が発見された．そこで，文頭切れ問題を解決するために，通話開始後，無音検出を契機として録音・音声認識によるテキスト化を確定するものの，録音・音声認識を停止せず，継続させる方式へ変更した．結果として，文頭切れの回避や音声認識精度の向上を実現している．

2.4　ユーザからの反応・ご意見

⑴機能改善

トライアルサービス提供期間中に，聴覚に障がいのある方の利用に適したサービス，アプリを実現するために，モニターユーザにアンケートを実施し，継続的な機能改善に努めた．ここでは，難聴かつ発話が難しい方々から多くの声をいただき，機能開発した「入力発話」について解説する．

⑷発話したい言葉を入力し，音声で伝える入力発話機能

通話中に専用アプリ（またはWebアプリ）から文字入力機能を起動し，発話したい言葉をテキスト入力・送信することで，合成音声による発話機能（通話相手に音声合成エンジンを介して音声再生する機能）を実装した［3］．合成音声は通話相手側だけでなく，サービス利用者側にも送話し，発話音声が重なった場合でもミキシングで通話音声を送話できるようにした．また，対話をスムーズにするため，あらかじめ，

サービス利用者が定型文を専用アプリ（またはWebアプリ）内に登録し，タップするだけで，通話中に簡単に発話ができる機能も実装した．

(b)難聴かつ発話が難しい方でも分かりやすいユーザインタフェース

通話相手の発話と入力発話機能を用いて発話した内容との前後関係を明確にすることで会話を成立させるため，文字入力したテキストも通話相手の発話とともにサービス利用者のスマートフォン上に表示可能にした．さらに，サービス利用者が相手の反応のタイミングを理解し，入力発話を適切なタイミングで行えるようにするため，合成音声の再生開始および終了タイミングをサービス利用者が正しく認識できるよう実装した．

実現方式としては，サービス利用者のスマートフォンとサービス処理装置間でWebSocket*3を利用し，合成音声の再生開始・完了タイミングで，サービス処理装置からスマートフォンへ信号を送信する機構を設け，合成音声の再生を開始した旨の信号を受信したタイミングで文字入力したテキストを表示し，合成音声の再生を完了した旨の信号を受信したタイミングで吹出しの色を変化させるように工夫した．

(2)アンケートによるユーザ評価の確認

トライアルサービスの目的であった，サービスコンセプトへの受容度，現行の認識精度での満足度について，ユーザアンケートを行い確認した．音声認識の誤変換は残るものの，これまでできないものと諦めていた音声通話が可能となることへの多くの期待と，継続して「みえる電話」サービスを提供することについて，多くのユーザからの支持を得たため，商用サービス提供を行うに至った．

3. 商用開発

3.1 概　要

トライアルサービスでは，専用の電話番号を利用する必要があることと，緊急呼，フリーダイヤル®*4などへの接続や，利用可能なサービスに制約があった．しかし，商用サービスでは，通常の090/080/070番号での利用を可能とし，緊急通報を含む音声通話サービスをサポートするサービスとして商用開発を行った（緊急通報対応のみ，サービス提供に向けて準備中）．また，トライアルではアプリ起動方式はSMS*5通知の受信を契機としていたが，プッシュ通知の受信契機へ変更した．

3.2 サービス実現方式

(1)システム開発

「みえる電話」商用サービスのシステム構成を**図3**に示す．音声呼処理は，サービスシナリオ実行基盤（vSCN：virtual Service Composition Node）*6とメディア処理装置（vMPN：virtual Media Processing Node*7）などで構成されるサービスイネーブラネットワーク（SEN：Service Enabler Network）*8基盤を利用して実現した［4］．音声認識エンジンおよび音声合成*9エンジンは，音声翻訳基盤内に格納し，既存サービスである，はなして翻訳とインタフェースの共通化を可能とした．

(a)呼処理

サービス利用者が発着信時，「みえる電話」サービスを利用するためにSEN基盤を経由して，通話相手との音声通話を接続する．IMS（IP Multimedia Subsystem）*10基盤においては，通常の通話音声はIMS基盤内のU-Plane転送装置（VGN，SIN）を介して接続されるが，「みえる電話」では，音声認識を行うため，通話音声を

＊3　WebSocket：Webサーバとクライアントの間でリアルタイム性がある双方向通信を実現するためのプロトコル．

＊4　フリーダイヤル®：NTTコミュニケーションズ㈱の登録商標．

＊5　SMS：テキストベースの短い文章を送受信するサービス．移動端末の制御用信号を送受信することにも用いられる．

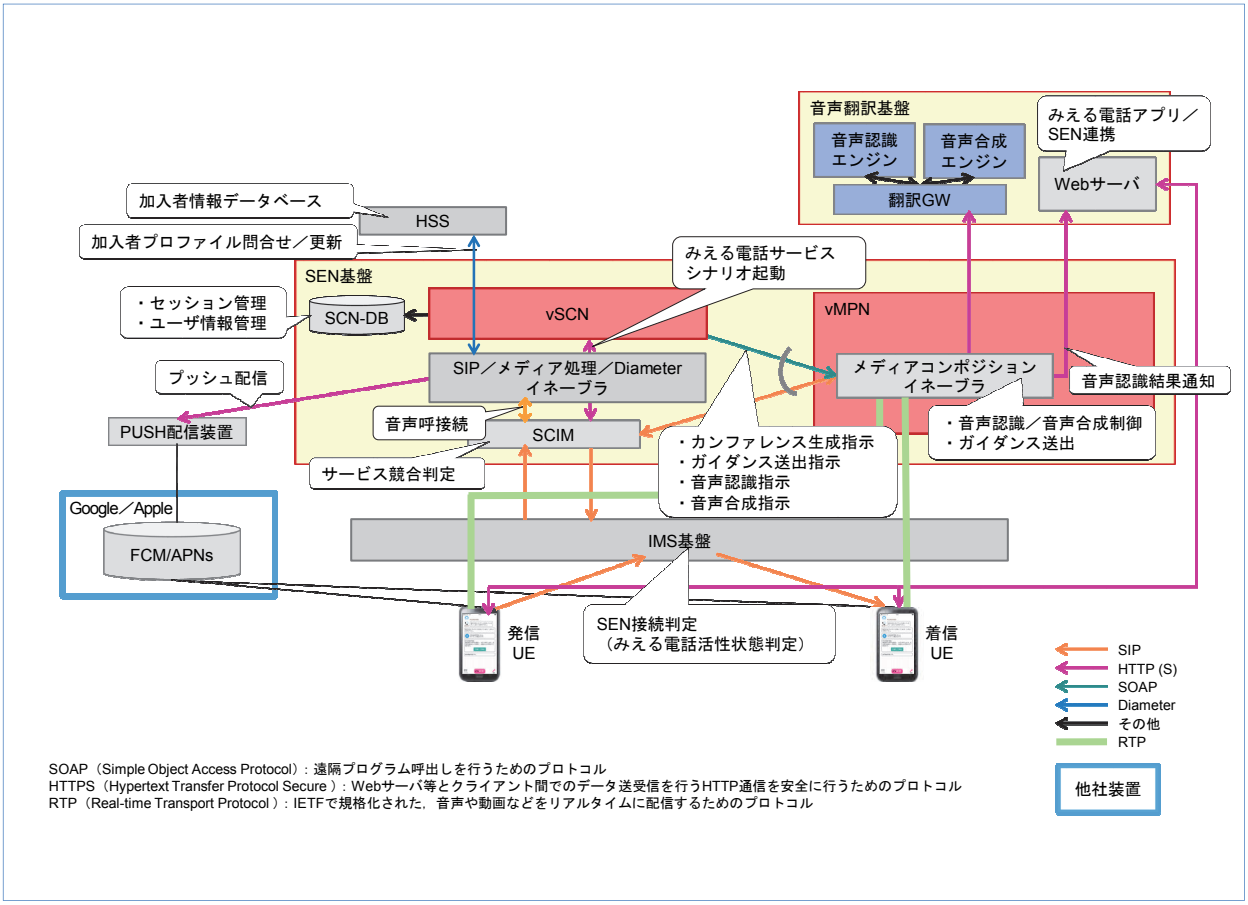
＊6　vSCN：サービスシナリオに基づきイネーブラ（＊18参照）を組合せてサービスを提供する装置．

＊7　vMPN：メディア処理装置．留守番電話やメロディコールといった音声メディアサービスなどさまざまなメディアサービスを提供している．

＊8　サービスイネーブラネットワーク（SEN）：複数のイネーブラ（＊18参照）を組み合わせることにより付加価値を提供する基盤．テレコム機能，Webアクセス機能，メディア制御などを具備する．

＊9　音声合成：テキストから人工的に音声データを作り出し，テキストを読み上げできるようにする技術．

＊10　IMS：3GPPで標準化された．固定・移動通信ネットワークなどの通信サービスをIP技術やインターネット電話で使われるプロトコルであるSIP（＊14参照）で統合し，マルチメディアサービスを実現させる呼制御通信方式．



SOAP（Simple Object Access Protocol）：遠隔プログラム呼出しを行うためのプロトコル
HTTPS（Hypertext Transfer Protocol Secure）：Webサーバ等とクライアント間でのデータ送受信を行うHTTP通信を安全に行うためのプロトコル
RTP（Real-time Transport Protocol）：IETFで規格化された，音声や動画などをリアルタイムに配信するためのプロトコル

図3 「みえる電話」システム構成

vMPNに引き込む処理が必要となる．具体的には，「みえる電話」サービス利用者が専用アプリにてサービス機能を有効にすることでIMS基盤（S-CSCF（Serving Call Session Control Function）*11）のSiFC（Shared initial Filter Criteria）*12情報にSCIM（Service Capability Interaction Manager）*13への接続情報が登録され，発着信に伴う音声呼接続要求を行うSIP（Session Initiation Protocol）*14-INVITE*15がSEN基盤のSIP受付部であるSCIMに送信されることになる．SCIMにおいて，加入者プロファイル*16情報により「みえる電話」契約判定および各種サービス間の競合判定を行い，vSCNの「みえる電話」シナリオを起動する．SIP／メディア処理／Diameter*17イネーブラ*18とサービスシナリオで，SCIMからのSIP-INVITEに基づき，vMPN内に会議室を生成し発着信者を参加者とするカンファレンスサービスを起動し，通話音声をvMPNに引き込む．

(b)ガイダンス制御／通話音声テキスト化

「みえる電話」サービスシナリオは，音声通話開始時にvMPNに対して，サービス案内のた

*11 S-CSCF：端末のセッション制御，およびユーザ認証を行うSIPサーバ．セッションとは，クライアントとサーバ，もしくはサーバ同士間でのやり取りされる一連の通信のこと．
*12 SiFC：要求信号をどのAS（サービスを提供するアプリケーション実行サーバ）に送信するかを判断するための基準，およびその機能．
*13 SCIM：ユーザからの要求に応じたサービスシナリオの選択やサービス競合の制御を行う機能．
*14 SIP：IETF（Internet Engineering Task Force）で策定された通信制御プロトコルの1つ．VoIPを用いたIP電話などで利用される．
*15 INVITE：SIPの信号の1つであり，接続要求を行うための信号．
*16 加入者プロファイル：契約，ユーザ設定，在圏情報などのサービス制御に必要な情報．
*17 Diameter：RADIUS（Remote Authentication Dial In User Service）をベースに機能を拡張したプロトコルでIMSにおける認証／認可／アカウンティングに利用される．
*18 イネーブラ：複数のサービスで使用できるように部品化された機能．

めの音声ガイダンスを送話する処理を指示した後，連続的に音声認識を行う処理を指示する．vMPNはサービスシナリオの指示に基づき，音声ガイダンスを送話後，音声認識処理を開始する．音声認識エンジンは入力された音声データを基に発話内容をテキスト化し，vMPNからテキスト化された発話内容をWebサーバ経由でサービス利用者のスマートフォンに送信する．スマートフォンで受信した通話音声テキストは，専用アプリ上で表示される．

連続的に通話音声テキストをアプリで表示するために，vMPNでは，通話中においても音声通話の無音区間[*19]を検知したタイミングで音声認識し続けられるよう，音声データを音声認識エンジンに送信し続けている．また，Webサーバとアプリ間は，Web Socketで接続することでリアルタイムに通話音声テキストを連続的にアプリで表示することを可能とした．

(c)入力発話機能

サービス利用者のアプリで文字入力されたテキストを，Webサーバから通知される音声合成要求に基づきvSCNがvMPNに対して，音声合成・合成音声再生を行う処理を指示する．vMPNはvSCNの指示に基づき，テキストを音声合成エンジンに送信し生成された合成音声を取得後，合成音声を再生する処理を開始する．合成音声は，通話音声にミキシングして，サービス利用者と通話相手に聞こえるように送話する．

(d)Android，iOS[*20]標準のプッシュ通知機能を利用したアプリ起動機能

通話時，フォアグランド[*21]に「みえる電話」アプリを表示するため，OS標準のPush通知機能を用いたアプリ起動方式を採用した．「みえる電話」アプリを起動するFCM（Firebase Cloud Messaging）[*22]/APNs（Apple Push Notification service）[*23]プッシュ通知の通知先を識別可能とするため，事前にFCMおよびAPNsから払い出されたデバイス識別情報をSEN基盤で保持しておき，通話開始時にサービス利用者のデバイス識別情報を載せたプッシュ通知要求をドコモ内GW[*24]であるPUSH配信装置[*25]へ送信する．プッシュ通知を利用したアプリ起動方式を採用することにより，メッセージアプリでSMSメッセージを受信し続ける必要がなく，アプリを起動することを可能とした．なお，iPhone[*26]では通知されたNotificationをタップすることでアプリが起動される．

(2)アプリ開発

トライアルサービスではAndroid端末のみ専用アプリを提供していたが，「みえる電話」商用サービスでは，Androidに加えてiOS向けの専用アプリも開発した．「みえる電話」アプリが通話中に表示する画面には，相手が発話した内容のテキスト変換結果が表示される．サービス利用者は聴覚に障がいがある方々であるため，表示された文章を読んで相手の発言を把握し，返答することで相手と通話することになる．「みえる電話」を利用しない通話に比べて，「文章への変換」と，「文章を読む」作業が追加されるため，そのままでは会話のテンポが遅くなる．

「みえる電話」では健聴者の通話と同程度のテンポで通話できるようテキスト変換結果の表示方法とUI[*27]に工夫を施している．変換結果の表示方法については，文章単位で認識が完了してからテキストを表示するのではなく，文字単位で途中変換結果をリアルタイムに表示し（**図4**），発話が完了したタイミングで文章全体を踏まえて修正した確定テキストを表示している（**図5**）．相手の発話に合わせて変換結果が文字単位で表示されるため，会話のテン

＊19 無音区間：通話回線上に通話音声が存在しないと判断される区間．

＊20 iOS：米国およびその他の国におけるCisco社の商標または登録商標であり，ライセンスに基づき使用されている．

＊21 フォアグランド：スマートフォンのホーム画面に他アプリの画面が表示されている場合でも，ユーザが直ぐに操作できるようにアプリなどを最前面に表示すること．

＊22 FCM：サーバからクライアントであるAndroid端末上のアプリにデータを送信できるようにするPUSH通知サービス．

＊23 APNs：PUSH技術を使って，常にオープンなIP接続を通してアプリのサーバからの通知をiPhone端末に転送するサービス．

＊24 GW：プロトコル変換やデータの中継機能などを有する関門装置．

＊25 PUSH配信装置：プッシュクライアントからSMS送信受付／応答を行う装置．

＊26 iPhone：Apple, Inc.の商標．ただし，日本国内ではアイホン株式会社のライセンスに基づき使用されている．

＊27 UI：ユーザとコンピュータとの間で情報をやり取りする際の操作画面や操作方法．

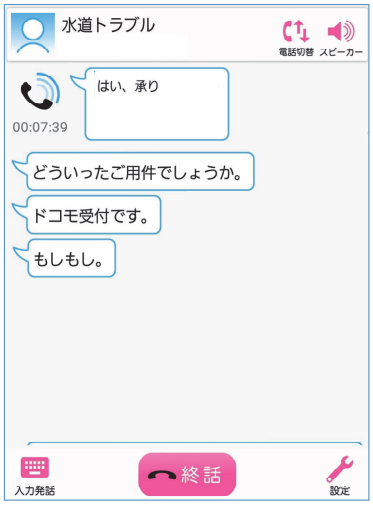


図4　通話中画面　途中結果表示

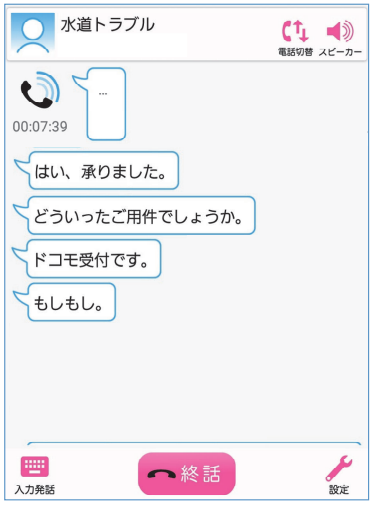


図5　通話中画面　最終結果表示

ポを極力損なうことなく通話することができる．また，相手が発話中であることがひと目で分かるように，音声の発話が始まった瞬間から画面上でアイコンを明滅させている．相手が「話しているのか」，「黙っているのか」がひと目で分かるため，サービス利用者自身が返答するタイミングがわかりやすくなっている．

4. あとがき

本稿では，「みえる電話」商用サービス化に向けた取組みと実現方式について，その詳細を解説した．聴覚に障がいのある方にとって生活をサポートする重要な役割を担うサービスであることはもちろん，周囲の騒音が大きく通話相手の声が聞き取りづらい環境など，利用シーンによっては健常者にとっても有効なサービスとなる可能性も秘めている．今後のサービス拡充に向けて，さらなる音声認識精度の向上に向けた検討を進める．

文　献

[1] 小磯 卓児，三上 和愛，佐藤 篤，太田 昌宏：“聴覚障がい者向け通話音声テキスト化サービスの実現方式検討，”電子情報通信学会2017年総合大会，2017.

[2] 三上 和愛，小磯 卓児，佐藤 篤，太田 昌宏：“通話音声テキスト化サービスの連続音声認識方法の改善，”電子情報通信学会2017年総合大会，2017.

[3] 小磯 卓児，三上 和愛，佐藤 篤，太田 昌宏：“聴覚障がい者向け通話音声テキスト化サービスへの入力発話機能の検討，”電子情報通信学会2017年ソサイエティ大会，2017.

[4] 飯村，ほか：“ネットワーククラウドを構成するサービスイネーブラネットワーク基盤（SEN）の導入，”本誌，Vol.20，No.2，pp.6-15，Jul. 2012.