一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2012-97(2012-12)

音声の到来方向により対象言語を切り替える自動通訳システム

辻川 剛範† 岡部 浩司† 花沢 健†

†日本電気株式会社 情報・メディアプロセッシング研究所
〒211-8666 神奈川県川崎市中原区下沼部 1753
E-mail: †tujikawa@cb.jp.nec.com, k-okabe@bx.jp.nec.com, k-hanazawa@cq.jp.nec.com

あらまし 音声の到来方向により対象言語を切り替える自動通訳システムを提案する．提案システムでは，二つのマイクを用いて，特定の方向から到来する音声を検出し，その音声の到来方向に応じて音声認識の対象言語を切り替える．これら音声検出と言語切り替えの自動化により，発話毎に要求される煩わしいボタン操作を省略することができる．対象言語の自動切り替えを含めた音声認識評価を行った．話者から 40cm 離れて位置する二つのマイクロホンで収録した発話に対して，提案システムでは約 80%の実用的な単語正解精度が得られることを確認した．

キーワード 自動通訳，音声認識，言語識別，到来方向，音声検出，マイクロホンアレイ

Automatic Speech Translation System Selecting Target Language by Direction of Arrival Information

Masanori TSUJIKAWA† Koji OKABE† and Ken HANAZAWA†

† Information and Media Processing Laboratories, NEC Corporation
1753, Shimonumabe, Nakahara-Ku, Kawasaki, Kanagawa, 211-8666, Japan
E-mail: †tujikawa@cb.jp.nec.com, k-okabe@bx.jp.nec.com, k-hanazawa@cq.jp.nec.com

Abstract An automatic speech translation system selecting target language by direction of arrival information is proposed. The proposed system uses two microphones to detect speech signals arrived from a certain direction. Depending on the direction of arrival information, target language for speech recognition is selected. Both the speech detection and the language selection release users from operations that are required for each utterance. In speech recognition evaluation for the proposed system, 80% of word accuracy was achieved for utterances recorded with two microphones, which were 40cm distant from speaker position.

Keyword automatic speech translation, speech recognition, language identification, direction of arrival, speech detection, microphone array

1. はじめに

我々は携帯電話等の端末上で動作するコンパクトな日英双方向自動通訳システムを開発してきた[1]．携帯電話端末は搭載されている画面が小さいため，ユーザ A が発声してから会話の相手のユーザ B がその翻訳結果を知るまでの下記手順 i から v に，長い時間を必要とする．その結果，会話のスムーズさが阻害されるという課題があった．

i. A が発声言語と入力開始の決定操作を行う．
ii. A が入力音声を発声する．
iii. A が認識結果と翻訳結果を確認する．
iv. A が B に携帯電話端末を渡す．
v. B が翻訳結果を見て内容を理解する．

近年，携帯電話端末に比べ大画面を搭載したタブレット端末が普及し始めている．このタブレット端末上で自動通訳システムを動作させれば，二人のユーザが同時に翻訳結果を閲覧しながら会話をすることができる．すなわちユーザ B は手順 iii と iv を待たずに内容を理解（手順 v）できる．結果としてユーザ A の発声からユーザ B の内容理解までに必要な時間が短縮され，会話のスムーズさの向上が期待できる．

しかしタブレット端末を用いるだけでは十分スムーズに会話をすることはできない．それは，発声の前に手順 i が依然として必要であり，発声の開始時に待ち時間が生じるためである．また手順 i は発声する度に必要であるため，操作自体も煩わしい．この手順 i を省くために，ボタン操作を伴わない高精度な音声検出技術[2][3]や言語識別技術[4]が求められる．さらに，ユーザ二人が同時に画面を閲覧するため，搭載マイクから数十ｃｍ離れた位置からの発声を高精度に認識で

きる，耐雑音音声認識技術が求められる．

本稿では，音声の到来方向により対象言語を切り替える自動通訳システムを提案する．提案システムでは，二つのマイクを用いた音声検出と検出した音声の到来方向に応じた言語識別を行う．またマイクから離れた位置での発声については，耐雑音処理として音声モデルを用いた雑音抑圧と音響モデルの雑音重畳学習を併用する[5]．以降では，提案手法，評価方法，およびその評価結果について詳述する．

2. 通訳タブレット

自動通訳アプリケーションを搭載したタブレット端末を，本稿では通訳タブレットと呼ぶ．ホテルや施設の受付ロビー，チケット販売窓口，オフィス，商業施設のカウンタなどで，母語の異なる二人の話者が向かい合い，端末を介して会話を行うことを想定している．使用例としては，二人の間にタブレット端末を寝かせて置き，お互いにその認識結果と翻訳結果を閲覧しながら会話をする使用例（**図 1**）と，二人の間から少し横にずらした位置にタブレット端末を立たせて置き，それを閲覧しながら会話をする使用例を想定している．

上記の実利用環境では背景雑音が存在し，音声認識精度を下げる一因となる．加えて，例えば隣のカウンタで同じように会話をする別の客と店員の音声（妨害音声）の存在も想定され，さらに音声認識精度を押し下げると考えられる．二人のユーザが通訳タブレットを同時に閲覧してスムーズな会話を行うためには，このような高雑音環境下において，高精度な音声検出，言語識別，音声認識を行う必要がある．

3. 提案手法

前述したような利用シーンを想定すると，通訳タブレットは次の 3 点の要件を満たす必要がある．

・高精度な耐雑音自動音声検出
・高精度な耐雑音自動言語識別
・高精度な耐雑音音声認識

音声検出と言語識別に関する要件に対して，二つのマイクを用いた音声検出（2 入力音声検出）を用いることを提案する．耐雑音音声認識については，音声モデルを用いた雑音抑圧と雑音重畳学習の併用により要件を満たす．本節では，これらの手法について説明する．

3.1. 2 入力音声検出と到来方向による言語識別

高精度な耐雑音自動音声検出と耐雑音自動言語識別を行うために，特定の方向から到来する音を検出す

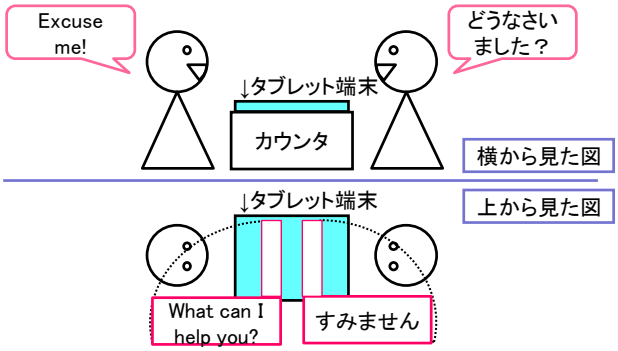

図 1 使用例

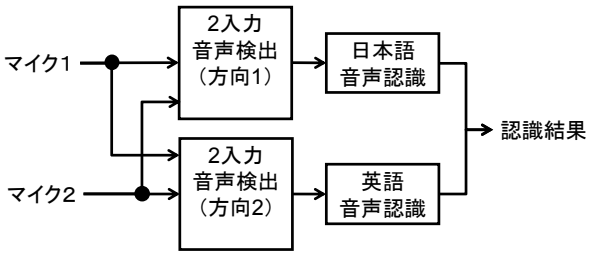

図 2 提案する音声の到来方向による対象言語の切り替え

る 2 入力音声検出[2]を用いた．本方式は，特定の方向から到来する音を除去するフィルタの出力と，その方向から到来する音を強調するフィルタの出力の比を用いて音声を検出する．出力の比が閾値より大きい場合にはその特定の方向から音声が到来していると判定し，逆に閾値より小さい場合にはその特定の方向からは音声が到来していないと判定する．本方式の特長は，複素スペクトル領域と振幅スペクトル領域の 2 段階で特定の方向から到来する音を除去することである．2 段階で除去することにより，音声の到来方向が想定からずれた場合にも頑健に音声を検出することができる．

この 2 入力音声検出を用いた提案手法の構成図が**図 2** である．端末から見て，ユーザ A の方向を方向 1，ユーザ B の方向を方向 2 として，各方向から到来する音声のみを検出する音声検出器を持つ．それぞれの方向以外からの音声を棄却することで，従来技術である 1 入力音声検出方式に比べて背景雑音や妨害音声による誤検出を減らすことが可能となる．

さらに，方向 1 と方向 2 がお互いに十分離れていれば，方向 1 の音声検出器は方向 2 からの音声も棄却することができる．この特性を利用し，方向 1 からの音声は日本語，方向 2 からの音声は英語，といったように到来方向を利用して言語識別を行うことができる．両言語の認識結果の尤度を比較する従来の言語識別手法[4]は必ずしも識別精度が高くないため，高精度な方向識別の結果を利用することで，より高精度な言語識

別が可能となる.

3.2. 音声モデルを用いた雑音抑圧と音響モデルの雑音重畳学習

マイクから数十 cm 離れた位置からの発声は，数 cm の場合と比較して入力音声の SNR（音声対雑音比）が低い. 高い SNR の音声データを用いて学習した音響モデルを用いて, 低い SNR の音声を認識すると認識精度が低下する．そこで，音声モデルを用いた雑音抑圧と音響モデルの雑音重畳学習とを行うことにより[5]，高精度な耐雑音音声認識の実現を狙った.

音響モデルの雑音重畳学習に使用した音声データは次の通りである．静かな環境で収録したクリーン音声データに対して, SNR が平均 5dB かつ標準偏差 3dB, また SNR が平均 15dB かつ標準偏差 3dB, また SNR が平均 25dB かつ標準偏差 3dB の正規分布になるように雑音データを重畳して雑音重畳音声データを作成した. この雑音重畳音声データによって学習した音響モデルとクリーン音声データで学習した音響モデルとを混合して雑音重畳音響モデルを作成した.

4. 評価音声の収録

提案手法の効果を確認するために，実際の使用環境を想定した評価実験を行った．本項では，評価用音声データの収録環境についての詳細を記述する.

7 インチタブレット端末の形状，大きさを備えるモックを用いて収録を行った. マイクは表側に 3cm 間隔で 2 個配置した．収録にはサイズが 5.0×5.0×2.2m, 残響時間 0.3s の防音室を用いた.

図 3 のように二つのマイクを搭載したタブレット端末モックと音声再生スピーカを配置した．静かな環境（騒音レベルは 28dBA）で，日本語音声と英語音声を交互にスピーカから再生し，端末モック上のマイクで収録した．θ は 0 度，15 度，30 度，45 度，60 度，L は 20cm，30cm，40cm でそれぞれ変化させた．再生した音声データは旅行会話読み上げ音声である．話者は日本語話者が 4 名, 英語話者が 4 名で, 日本語話者, 英語話者が 1 発声ずつ交互にそれぞれ 5 発声, 計 10 発声した後，日英両方の話者を交代させ，計 40 発声を再生して収録した.

さらに，遠方に配置した別のスピーカで音を拡散させるように（室内の壁に向けて）雑音を再生し，前記の旅行会話音声と同様にタブレット端末モックに搭載された 2 つのマイクで，背景雑音のみを収録した. 背景雑音として再生した雑音データは，オフィス雑音(50dBA)，ロビー雑音(45dBA)，窓口雑音(57dBA)の 3 種類である．このように収録した背景雑音を，前述した音声データに人工的に重畳して評価用音声データを作成した.

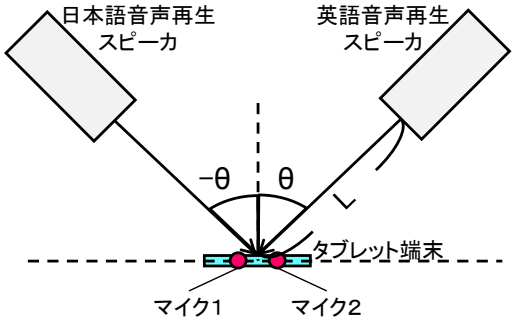

図 3 マイクとスピーカの配置

表 1 音声対雑音比および音声対妨害音声比 [dB]

	距離	20cm	30cm	40cm
雑音環境	静か	40.4	37.0	35.0
	オフィス	21.2	17.8	15.8
	ロビー	21.1	17.7	15.7
	窓口	12.6	9.1	7.1
	妨害音声	14.6	11.2	9.2

また，店頭のカウンタ等での使用時に，隣のカウンタで別の客と店員の会話が行われていることを想定し, その音声を妨害音声と定義し，次のように収録した. 妨害音声は θ =45 度, L=30cm, かつタブレット端末モックと日英各音声再生スピーカの垂直方向距離を 1m にして，音声を再生し，収録した．この妨害音声データを前記の評価データにさらに人工的に重畳することで，妨害音声ありの評価データを作成した．音声対雑音比および音声対妨害音声比は**表 1** の通りである.

5. 評価実験

5.1. 音声検出評価

前項で述べた音声，雑音および妨害音声データを重畳して作成した評価データを用いて，2 入力音声検出の評価を行った．評価指標には，以下の式で定義する検出率と棄却率を用いた.

$$\text{検出率}[\%] = N_{utt}(b \cap c)/N_{utt}(a)\times 100$$
$$\text{棄却率}[\%] = N_{utt}(b \cap d)/N_{utt}(a)\times 100$$

$N_{utt}(x)$は x の条件を満たす発話の数である. ここで, a から d の条件は，以下である.

a.全発話（40 発話）
b.無音区間全体の 90%以上を無音と判定した発話
c.音声区間全体の 90%以上を音声と判定した発話
d.音声区間全体の 90%以上を無音と判定した発話

今回の 2 入力音声検出方式では，特定の方向から到来する音声を検出するため，当該特定方向の検出率は高く，棄却率は低いことが望ましい．一方，当該特定方向以外の検出率は低く，棄却率は高いことが望ましい．

図 4 は妨害音声なし，**図 5** は妨害音声ありの結果であり，3 種類の背景雑音での評価結果の平均を表している．図の左側には-45 度，右側には 45 度の方向から到来する音声を検出するように設定した場合の検出率と棄却率を表す．また音声を再生したスピーカとマイクの距離 L=30cm の条件の結果を示している．**図 4** の-45 度において，-45 度と-60 度から音声が到来する場合の検出率が 85%以上，棄却率が 0%という結果である．また 60 度から-15 度の方向から音声が到来する場合の検出率が 0%，棄却率が 95%以上という結果である．期待通り，-45 度に近い方向から到来する音声を高精度に検出し，それ以外の方向からの音声は高精度に棄却できている．また**図 4** と**図 5** の差は小さく，2 入力音声検出において-45 度から到来する音声を検出するように設定した場合には，妨害音声を棄却できている．以上のことは 45 度についても同様のことが言える．

図 4 と**図 5** の結果は，日本語音声の到来方向を-45 度，英語音声の到来方向を 45 度と事前に決めておけば，背景雑音や妨害音声が存在する環境で，話者の口とマイクの距離が 30cm 程度離れた場合においても，2 入力音声検出によって高精度な自動音声検出および到来方向による高精度な自動言語識別が可能であることを示している．

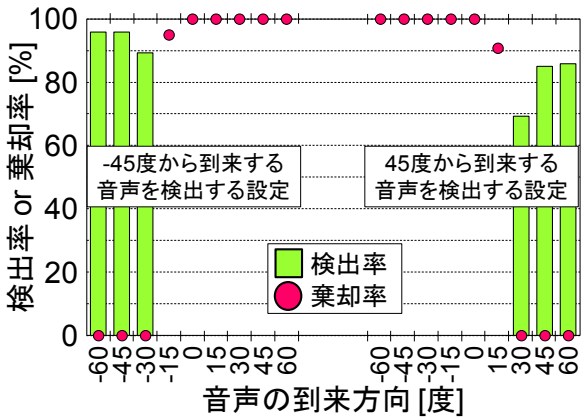

図 4 音声検出評価（妨害音声なし，L=30cm）

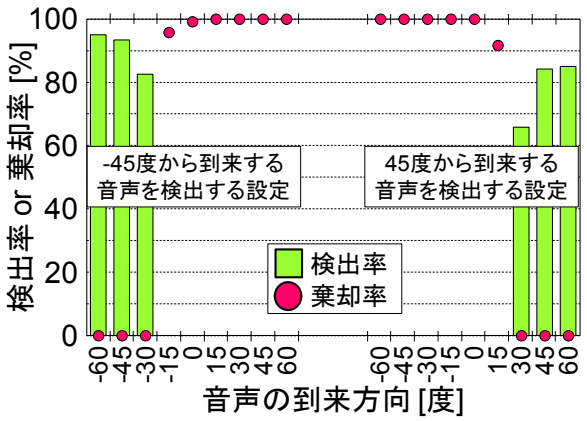

図 5 音声検出評価（妨害音声あり，L=30cm）

5.2. 音声認識評価

次に，音声認識実験による評価を行った．θ =45 度，L=20，30，40cm，背景雑音 3 種類，妨害雑音の有無という条件を変更させながら，**表 2** の A，B，C の方式を用いて音声認識実験を行い，それらの認識精度を比較した．

言語識別誤りを起こした場合，発声した言語に関しては脱落誤り，認識結果として出力された言語に関しては挿入誤りをそれぞれ起こしたとみなして単語正解精度（Word Accuracy）を算出した．

3 種類の背景雑音での結果を平均し，妨害音声なしの結果を**図 6** に，妨害音声ありの結果を**図 7** にそれぞれ示す．

ベースラインである A と比較して，2 入力音声検出を用いた B では，認識精度が大きく改善した．平均誤り削減率が，妨害音声なしでは 44.0%，妨害音声ありでは 70.2%である．背景雑音や妨害音声を音声区間と誤検出したことによる挿入誤りが，2 入力音声検出により抑制できたためである．

B と比較して，言語識別に 2 入力音声検出で検出した音声の到来方向を用いた C では，平均誤り削減率が英語で 5.1%，日本語で 3.7%であり，その効果を確認することができた．尤度基準の言語識別では日本語を英語に誤識別していたものが多かったが，到来方向を用いることで高精度に言語を識別できるようになり，英語での挿入誤りが改善した．

C の条件に注目すると，妨害音声ありの条件で，日本語，英語ともマイクとスピーカが 40cm の距離で約 80%の単語正解精度が得られている．提案システムにより，会話のスムーズさを向上しつつ，実用レベルの精度が確保できている．

6. まとめ

本稿では，よりスムーズに会話を行うことができる通訳タブレットを実現するために，音声の到来方向に

より対象言語を切り替える自動通訳システムを提案した．提案システムでは，2 入力音声検出による音声検出と言語切り替えの自動化により，発話毎に要求される煩わしいボタン操作を省略することができる．対象言語の自動切り替えを含めた音声認識評価を行い，話者から 40cm 離れて位置する二つのマイクロホンで収録した発話に対して，提案システムでは約 80%の実用的な単語正解精度が得られることを確認した．

文　献

[1] 花沢健，奥村明俊，岡部浩司，安藤真一，“高速・高精度なコンパクト・スケーラブル自動通訳ソフトウェアの開発と実用性評価，”情報処理学会論文誌 コンシューマ・デバイス＆システム（CDS），vol.2, no.2, pp.10-18, Jul.2012.

[2] 辻川剛範，“ハンズフリー音声認識のための 2 マイクロホンによる頑健な音声区間検出法，”2005 年春季日本音響学会講演論文集, pp.121-122, Mar. 2005.

[3] 荒木章子，藤本雅清，石塚健太郎，澤田宏，牧野昭二，“音声区間検出と方向情報を用いた会議音声話者識別システムとその評価，”2008 年春季日本音響学会講演論文集, pp.1-4, Mar. 2008.

[4] M. A. Zissman, “Comparison of Four Approaches to Automatic Language Identification of Telephone Speech, ”IEEE Transactions on Speech and Audio Processing, vol.4, no.1, pp.31-44, Jan.1996.

[5] 辻川剛範，荒川隆行，磯谷亮輔，服部浩明，“Model-Based Wiener Filter と Multi-Condition 学習の併用による車内音声認識，”2008 年春季日本音響学会講演論文集, pp.179-182, Mar. 2008.

表 2 比較する方式

	A	B	C
音声検出	1 入力	2 入力	2 入力
言語識別	並列サーチ 尤度基準	並列サーチ 尤度基準	音声の 到来方向

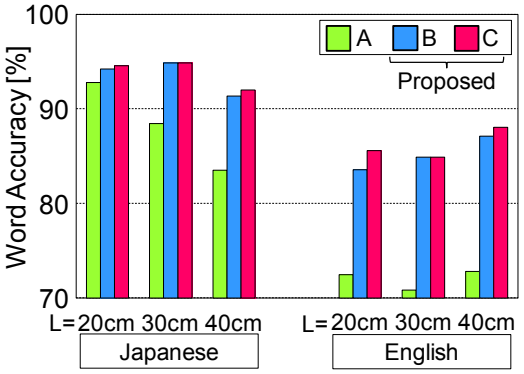

図 6 音声認識評価（妨害音声なし）

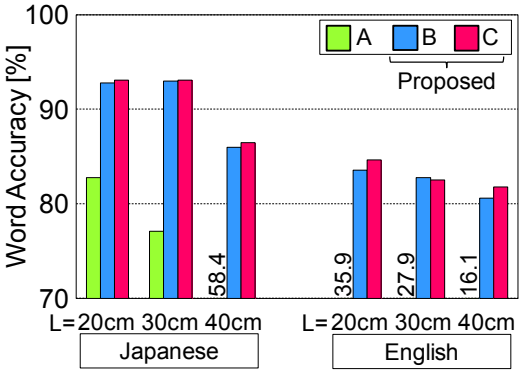

図 7 音声認識評価（妨害音声あり）