機器の遠隔操作を実現する音声インタフェース

Voice Interface for Operation of Distant Equipment

大内　一成　　古賀　敏之
■OUCHI Kazushige　■KOGA Toshiyuki

離れた場所の機器を音声認識で操作する場合，音声認識の開始を機器に伝えたり，周囲の雑音の影響を抑えたりすることが，実用的な認識性能を確保するうえで重要である。

東芝は，音声認識の前処理として，マイクロホンを２個用いて目的方向の音を強調し，それ以外の方向からの音を抑圧するマイクロホンアレー技術を活用し，離れた場所からでも実用的な精度で音声認識による操作が可能な音声インタフェースを開発した。拍手を合図に，それと同時に拍手音の到来方向を推定しマイクロホンの指向性を拍手音（話者）の方向に設定することで，話者の音声だけを強調して入力する。テレビの音声操作をモチーフにした性能評価で，4.5ｍの距離から実用的な性能で音声認識による操作が可能なことを確認した。

In order to operate distant equipment by a speech recognition system, there are two technical challenges for the realization of practical recognition accuracy: (1) commanding the equipment to start speech recognition, and (2) reducing the influence of ambient noises.

Toshiba has developed a voice interface for the operation of distant equipment that utilizes a microphone array technology to emphasize the sound in the target direction and suppress noises from other directions. When a user activates the speech recognition system by clapping twice, the system simultaneously detects the direction of the clapping and sets the directivity angle of the microphones to that direction so as to prioritize the input of the target user's voice. We have conducted evaluation experiments using the operation of a TV set as a motif, and confirmed that users can operate a TV from 4.5 m away by means of speech recognition with a practical level of performance.

1　まえがき

リモコンは，対象の機器を離れたところから操作可能にするものとして1950年代に登場してから現在まで，家庭やオフィスなど様々な場所で日常的に用いられている。近年，機器の高機能化が進み，特にテレビやオーディオなどのAV機器では本体で全ての操作を行うことが困難になってきており，リモコンの使用を前提にした商品仕様になっている機器も多い。

リモコンによる操作は，機器から離れた場所から所望の操作ができ非常に便利である一方で，次のような課題もある。

(1)　手元にないと取りに行く（探す）必要がある
(2)　ボタンが多すぎてわかりにくい
(3)　文字入力がたいへんである
(4)　機器専用のリモコンが部屋中にたくさんある

これらを解決する手段として，機器を人と接するように使用できる，音声による機器操作が期待されている。しかし，人が発する音声を点音源と仮定すると，その強度は距離の２乗に反比例して減衰するため，音声認識を精度よく行うには，マイクロホンと話者の口元との距離をできるだけ近づけることが望ましい。このような試みとして，テレビの番組検索にリモコン内蔵のマイクロホンを用いて検索キーワードを音声で入力するシステムがある[1]。前述の課題(3)に対し，音声認識による文字入力により従来に比べて大幅な操作時間の削減を実現した。しかし，電源のオン，オフやチャンネル切替えなどの単純な操作には，リモコンでボタン操作したほうが音声で入力するより早くて確実である。

リモコンが手元にない場合に機器操作を音声で行う例としては，東芝のエアコン用ボイスコントローラ[2]がある。これは，エアコン（並びにテレビと照明）のオン，オフや，温度調節などを音声で操作できるが，距離の制約からボイスコントローラ本体をユーザーから１ｍ以内に置くことを推奨している。一方，ボイスコントローラを経由せず機器本体に設置したマイクロホンで音声認識させる場合，エアコンやテレビではユーザーとの距離が遠くなり，音声の減衰や周囲雑音の影響で実用的な音声認識性能の確保は困難であった。

そこで当社は，音声認識の前処理として，２個のマイクロホンで目的方向の音を強調し，それ以外の方向からの音を抑圧するマイクロホンアレー技術[3, 4]を活用して，機器から離れた場所からでも実用的な性能で音声認識による操作ができる音声インタフェースを開発した。

2　遠隔からの音声認識入力の課題

離れた場所の機器を音声認識で操作する場合，実用的な

認識性能を確保するには次のような課題が挙げられる。

(1) 入力を意図しない音声による誤認識の抑制

(2) 話者の方向以外からの雑音による誤認識の抑制

(1)に対しては，従来，ボタンを押下して明示的に音声認識の開始を指示した後目的の音声を入力する，主にボタン操作による音声認識入力の開始指示が使われてきた。これに対し，当社のエアコン用ボイスコントローラでは，ボタン操作以外に拍手により音声認識入力の開始を指示できるため，コントローラから離れた場所からの開始指示と音声認識による操作が可能である。

(2)に対しては，2個のマイクロホンで話者方向の音を強調し，それ以外の方向からの音を抑圧するマイクロホンアレー技術を音声認識の前処理に用いることで，話者の方向に限定して音声を選択的に入力し，他人の話し声や環境雑音を大きく抑圧することが期待できる。また，話者の発話以外に重畳されて入力される周囲雑音や，機器からの音（テレビ番組の音声やエアコンの動作音）を抑えるノイズキャンセル技術やエコーキャンセル技術[5]の使用も効果が期待できる。

3 開発した音声インタフェースの概要

今回開発した手法の処理の流れを**図1**に示す。2回の拍手により音声認識の開始を機器に指示すると同時に，拍手音の方向を推定する。マイクロホンアレー技術により，マイクロホンの指向性を拍手音の方向にソフトウェアで設定することで，操作意図を持つユーザーの声だけを強調して抽出することができる。ここでは，拍手検出処理と拍手方向推定の処理について述べる。

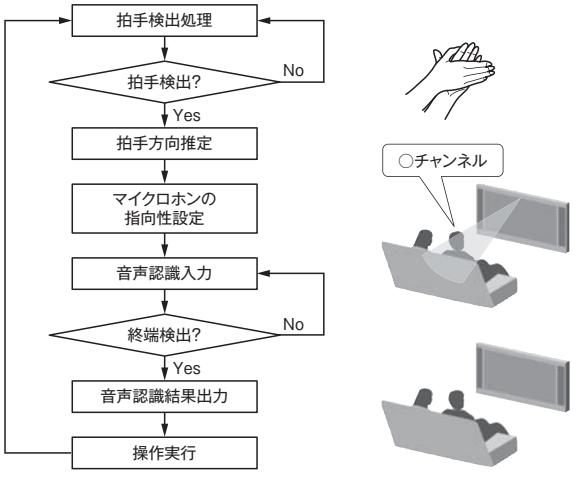

図1．提案手法の処理概要 ― 2回の拍手で音声認識を起動すると同時にマイクロホンの指向性を話者方向に設定し，その話者の声を強調して音声認識する。

Overview of voice interface processing

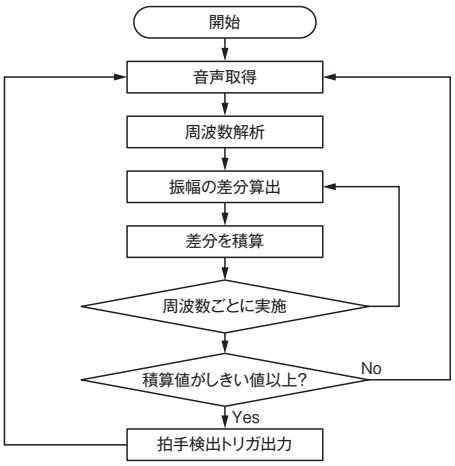

図2．拍手検出処理 ― 音声帯域よりも高い周波数帯域の成分も加味して拍手を検出する。

Flow of clap detection processing

3.1 拍手検出処理

音声認識入力の開始指示として拍手を2回行うこととした。拍手を1回にすると，ドアの開閉や物を落とした音などで誤検出の可能性がある。また，拍手を2回にするのは，相手に何かを伝えたり，依頼したりする場合などで慣れ親しんでおり，操作開始の合図として適していると考えた。

拍手検出処理のフローを**図2**に示す。一般に，人の音声は1 kHz以下であるのに対し，拍手の音は1 kHz以上も含め広い周波数を含んでいる。そこで，大きな声などによる誤検出を抑えるため，音声を取得した後，高速フーリエ変換（FFT）で周波数解析を行い，8 kHzまで周波数別に現在の振幅値と1サンプル前の振幅値の差分を取る。拍手音による急しゅんな立上がりを検出するため，この差分値が正（音量レベルが増加）の場合にその二乗値をスコアとして積算し，全周波数帯域でそのスコアを評価して拍手検出を行うこととした。

3.2 拍手方向推定

マイクロホンアレーを用いて音源方向を推定する手法としては，到来時間差（遅延和）法や，MUSIC（Multiple Signal Classification）法，ビームフォーマ法などがある。これらの処理は主に無線通信分野で電波の到来方向などを算出するために用いられるが，扱う周波数帯を変えるなどして音響処理分野に応用されている。ここでは，ある瞬間に発生した音が位置の異なる複数のマイクロホンに到達する時刻の差を求め，その差の集計から音源の方向を推定する到来時間差法を活用して拍手方向推定を行った。

図3に示した2個のマイクロホンA，Bに正面方向及び左45°方向から拍手した際の，拍手音のマイクロホンへの入力波形の例を**図4**に示す。拍手音入力のサンプリング周波数は48 kHzとした。

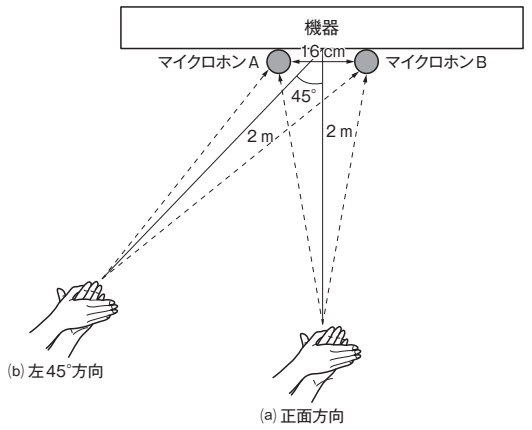

図3．拍手の位置とマイクロホンアレーの関係 — 斜めの位置からの拍手音は，2個のマイクロホンに時間差を持って入力される。
Relationship between clapping position and microphones

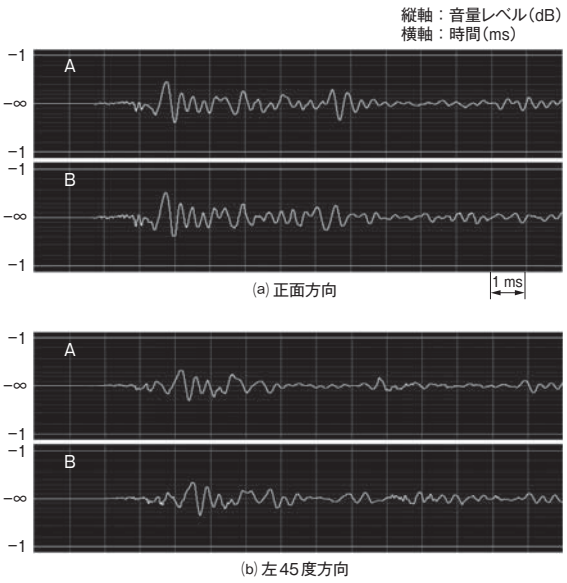

図4．拍手音の入力例 — 正面からの拍手音は左右のマイクロホンに同時に入力されるが，左方向からの音は先に左側のマイクロホンAに，遅れて右側のマイクロホンBに時間差を持って入力される。
Examples of waveforms of clapping sound

正面方向からの拍手音の波形は，マイクロホンA，Bに同時に音が入力されており，左45°方向からの波形は，まず距離が近いA，次に数百μsほど遅れてBに入力される。二つのマイクロホンへの入力波形は，元が同じ音源であるのでよく似た波形であり，この時間差を求める方法として，二つの波形の正規相互相関（NCC：Normalized Cross Correlation）を用いた。取得した音声の最新の1,920サンプル（40 ms分のデータ）の波形に対し，その区間において前後最大48サンプルずれた波形の相関値をそれぞれ算出し，その相関値が最大となるずれ幅D（サンプル）は式(1)から求められる。

$$D = \frac{d1 \cdot f}{V} \sin\left(\frac{\pi\theta}{180}\right) \qquad (1)$$

ここで，$d1$はマイク間距離，fはサンプリング周波数，Vは音速であり，θは音源方向角度で定義域は±90°とする。

4　実験による性能評価

4.1　基本性能

拍手検出及び拍手方向推定の基本性能の評価用データとして，7名の被験者が拍手2回を，図3に示す距離関係で，正面，右45°，左45°の各方向から計10回ずつ行い，その音を16ビット，48 kHzで収録した。また，操作対象機器としてはテレビを想定し，テレビがオン状態とオフ状態のそれぞれで拍手音を収録した。

被験者が拍手2回を連続10回行い，3.1節で述べたスコアを求めた結果を**図5**に示す。ここで，スコアを$S(t)$（t：時間）と記す。$S(t)$＝1,000をしきい値とすることで，収録した拍手2回は全て検出できることを確認した。またテレビ番組の音による誤検出も観測されず，開発した手法によって精度よく拍手を検出できることがわかった。その後，日常生活の環境下での雑音データを約36時間分収録して誤検出頻度の評価を実施したところ，マイクロホンの近くで子供がボール遊びをした際に2回，物を床に落としてバウンドし連続して大きな音が2回発生した際に1回の誤検出が発生した。今回開発した手法のように音情報だけでこれらを排除することは困難であるが，その後の音声認識処理により排除することで，誤操作の削減が期待される。

拍手方向推定の結果を**表1**に示す。方向推定結果と正解との許容誤差を±1°とすると，50％未満の正解率となる場合がある。マイクロホンの指向性設定幅は実用上±10°あるいはそれ以上とすることが望ましいため，許容誤差範囲を±10°と

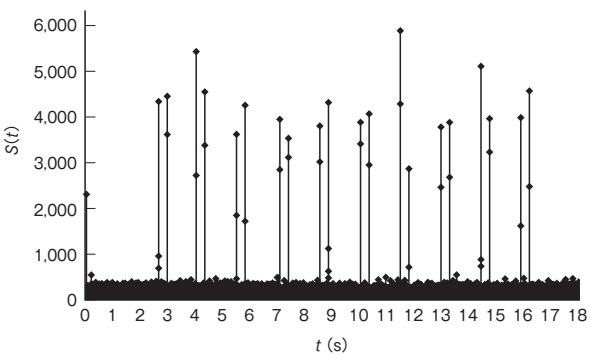

図5．拍手検出結果の例 — $S(t)$＝1,000をしきい値とすることで，拍手2回を全て検出できることを確認した。
Example of results of clap detection

表1. 拍手方向推定結果
Results of clap-detection estimation

拍手方向		テレビの状態	誤差許容範囲			
			±1°	±5°	±10°	±12°
正解率 (%)	正面	オン	53.6	97.9	99.3	100.0
		オフ	46.0	93.0	98.0	99.0
	左45°	オン	49.3	96.4	100.0	100.0
		オフ	51.0	100.0	100.0	100.0
	右45°	オン	31.4	83.6	95.0	99.3
		オフ	50.0	90.0	94.0	100.0

すれば拍手方向推定の正解率は94.0％以上の，±12°とすれば99.0％以上の精度でそれぞれ拍手音の方向を推定できることを確認した。

4.2 実用性能

テレビでの利用を想定し，拍手2回による音声認識の起動から音声認識による目的操作の達成までの実用性能を実験により評価した。国内におけるテレビの視聴距離は，画面サイズにもよるが住宅事情もあり2.5 m程度で飽和している[6]。一方，米国の平均視聴距離は3.37 mとの報告[7]がある。リビングを模した実験室で被験者5名に対して，これらを勘案して，近距離を1.5 m，遠距離を4.5 mとした二つの視聴距離で，電源オン，オフや，チャンネル切替え，音量調節などの20タスクを提示し，拍手検出率，方向推定正解率，及び音声認識率の各性能を評価した。

音声認識には，チャンネル名や主要操作など60語が登録された孤立単語認識エンジンを用いた。音声認識に用いる音響信号のサンプリング周波数は16 kHz，フレーム幅は25 ms，シフト幅は8 msで分析した。音響特徴量には0次から12次のMFCC（Mel-Frequency Cepstral Coefficient）とその1次回帰係数（Δケプストラム）及び2次回帰係数（ΔΔケプストラム）で構成される計39次元の特徴ベクトルを用いた。前処理にはノイズキャンセルとエコーキャンセルは適用せず，音声認識開始後はテレビ音声をミュートにした。音響モデルには3状態20混合のleft-to-right型の隠れマルコフモデル（HMM：Hidden Markov Model）を用いた。

拍手検出率，方向推定正解率，及び音声認識率の結果を距離別に**表2**に示す。ここで，マイクロホンの指向性は目的方向に対して±15°に設定したため，方向推定正解率は推定結果が±12°以内にあれば正解として扱った。

成功率は，4.5 mの場合は1.5 mの場合に比べて若干下がるが，遜色ない成功率が実現できており，今回開発した手法の妥当性を客観的に示す結果となった。また，テレビとの距離が遠くなると，遠くの人に話す際に声を大きくするように，拍手のたたき方や音声認識入力時の声の大きさが，自然に大きくなる傾向があることが確認できた。拍手2回から音声認識による目的操作指示まで，約80％以上の場合は一度も失敗せずタスクを達成できた。

表2. 実用性能評価結果
Results of practical evaluation at distances of 1.5 m and 4.5 m

評価項目		視聴距離	
		1.5 m	4.5 m
拍手検出率	(%)	92.0	91.0
方向推定正解率	(%)	100.0	100.0
音声認識率	(%)	94.0	92.0

5 あとがき

対象となる機器を，リモコンによらず拍手と音声だけで操作可能な音声インタフェースを開発した。音声認識開始の合図を拍手2回とし，その拍手音の方向にマイクロホンの指向性を設定することで，周囲の雑音の影響を軽減させ，離れた場所からでも精度よく音声認識による操作を可能にした。

機器の高機能化に伴い，使いやすさが犠牲になってはならない。今後も，人と接するように，誰でも使える，自然なユーザーインタフェースの実現に向けた研究開発を進めていく。

文 献

(1) 大内一成．日常的使用を目指した音声認識検索システム．東芝レビュー．**65**，5，2010，p.64－65.
(2) “東芝エアコン用ボイスコントローラ VOiPY［ボイピィ］形名：RB-VC01”．東芝ホームページ．<http://www.toshiba.co.jp/living/air_conditioners/pickup/rb_vc01/>，（参照2013-08-09).
(3) 天田 皇 他．音声認識のためのマイクロホンアレー技術．東芝レビュー．**59**，9，2004，p.42－44.
(4) 天田 皇．車載向け学習型マイクロホンアレー技術．東芝レビュー．**64**，2，2009，p.35－38.
(5) 紀伊雅之 他．携帯機器の高音質化を実現するスピーカアンプLSI TC94B-23WBGと音声信号処理コーデックLSI TC94B24WBG．東芝レビュー．**67**，10，2012，p.21－24.
(6) 窪田 悟 他．家庭におけるテレビの観視条件．映像情報メディア学会誌．**60**，4，2006，p.597－603.
(7) Nathan, J. G. et al. Television Viewing at Home：Distances and Visual Angles of Children and Adults. Human Factors. **27**, 4, 1985, p.467－476.

大内 一成 OUCHI Kazushige
研究開発センター インタラクティブメディアラボラトリー主任研究員。状況認識技術とそれを活用した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，人間情報学会会員。
Interactive Media Lab.

古賀 敏之 KOGA Toshiyuki
デジタルプロダクツ＆サービス社 ビジネスソリューション事業部設計第六部主務。デジタルプロダクツのソフトウェア開発に従事。
Business Solutions Div.