音声認識による AV 機器向け音声対話インタフェース
Spoken Dialog Interface for Audio Visual Equipments by Using Speech Recognition

高原　健悟
Kengo Takahara
法政大学情報科学部ディジタルメディア学科
E-mail: kengo.takahara.tu@cis.hosei.ac.jp

Abstract

An objective of this thesis is to utilize for tuning TV and for searching tunes or artists in a music player like iPod by building system which can search heads using speech recognition. At first, as the most important part of this system is speech recognition, it is necessary to examine how many words a software of speech recognition can recognize words correctly. In this thesis, utterances which were recorded in the silent place and in the noisy place such as in the car or in the street along a causeway, which is near to an effective surrounding, were used, and the difference of each recognition result was deliberated. And regarding recognition rate, it was classified by speakers and words in order to examine if the difference of recognition rate depends on speakers or if there are low rate words which don't depend on speakers and was deliberated. A head set microphone was used to record utterances and evaluated its performance if it is eligible recording equipment for this system. In the present step, total recognition rate of the utterances which were recorded in the silent place is about 85%, and in the noisy place is about 77%. A line hereafter is to improve this rate and to build system which can search heads using speech recognition.

1. まえがき

AV 機器の利便性は向上の一途を辿り，その中で音声認識を利用した機器操作の技術も導入されてきた．携帯電話やカーナビゲーションシステムがその代表例である．音声による機器操作をリモコン，キーボード，マウスなどのインタフェースを使用した場合の機器操作と比較すると，キーボードやマウスなどが不要となり，機器やコンピュータ操作を詳しく知らない操作者でも容易に利用可能であるという利点が期待できる[1]．しかし，実際に音声認識技術を利用したこれらの商品を，私たちの日常生活の中で見かける場面は少ないのが現状である．信頼性が必要とされる商品の分野への本格的な導入には，音声の認識精度の向上，操作の確実性向上，操作インタフェースの開発など解決すべき課題が多いことがその理由と言える[1].

その中で本研究では，音声の認識精度や集音機器の性能を実験によって考察する．またそれと同時に，入力音声を認識し，その認識結果に対応する項目を検索し，検索結果をリプライするという方法で機器操作を行うような機能をテレビや音楽再生機器などの AV 機器に搭載することで，その利便性は更に向上すると考える．ゆえに本研究の目的は，音声認識により項目検索を行えるシステムを作成し，そのシステムを利用することにより，テレビのチャンネル選局や，音楽再生機器においての曲名やアーティスト検索に活用していくことである．

2. AV 機器向け音声対話インタフェース

AV 機器向け音声対話インタフェースに関して，主に音楽再生機器における曲名やアーティスト名検索とテレビのチャンネル選局を考える．

音楽再生機器における曲名やアーティスト検索に関しては，i-Pod などの音楽再生機器に向けて曲名やアーティスト名を発話することで，その曲やアーティストを自動的に検索し表示するシステムを作成する．

また，テレビのチャンネル選局に関しては，テレビのリモコン，あるいはテレビ本体に向けて「NHK」や「1 チャンネル」などといった言葉を発話し，それを認識することでテレビのチャンネルを自動的に切り替えるシステムを作成する．それぞれの音声対話の例を図 1 に示す．

いずれのシステムの場合についても音声を集音し認識する必要があるので，まずは集音に使用するマイクを決定しなければならない．本研究では，小型であることや雑音が比較的入りにくいことなどを考慮し，ヘッドセット型マイクを用い収録を行う．

このマイクで収録した音声に関しては，音声認識ソフトでどの程度正しく認識できるかということも評価する必要がある．そのため，多数の音声サンプルをそのマイクで収録し，音声認識ソフトで認識させ，認識率を求めることが必要となる．性別，滑舌の良し悪し，声質などの違いから，人や単語によって認識率が異なってくることも考えられるので，出来るだけ多くの種類の音声を収録し，人別，単語別に認識率を求める．

音楽再生機器に関しては，ほとんどの場合，周りに騒音がある場面が想定されるため，そのような状況下でも認識率を低下させないようにするという点が課題となる．また，テレビのチャンネル選局に関しては，誤ってテレビから発せられた音声を認識してしまうことが考えられる．そのため，どのように人の声とテレビの音声を認識し分けるかという点が課題となる．

また，音声認識技術を使用して AV 機器を操作しようとする際，ユーザが命令語となる単語列を覚えなくてはならないという手間を要する．これも課題の 1 つだが，ユーザが「単語列を覚える」という作業をせず，普段の言葉で自由に発話できることを目指した家電機器操作システムの提案を行っている文献もある[2].

いずれの場合においても，操作を行うユーザ本人のみの音声を認識できるようにすることが目標である．

Supervisor: Prof. Katsunobu Ito

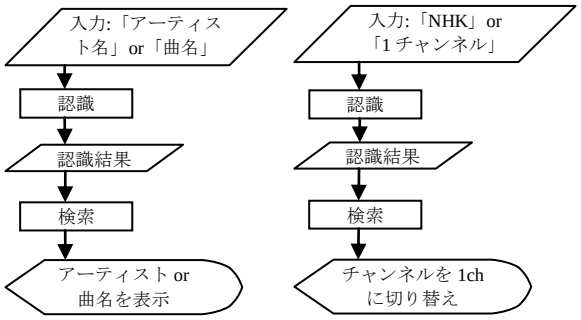

図 1　音声対話の例

3. 音声認識

3.1. 音声認識ソフトウェア Julian

音声認識には，フリーの高性能音声認識ソフトウェア Julian を使用する．Julian は数万語の語彙を対象とした文章発声の認識を行う能力を持ち，発音辞書や言語モデル･音響モデルなどの音声認識の各モジュールを組み替えることで，様々な幅広い用途に応用できる．Julian を用いた認識システムの構成を図 2 に示す．言語モデルとして有限状態文法を，音響モデルとして HMM を使用し，入力を 2 回に分けて処理する 2 パス探索を行う

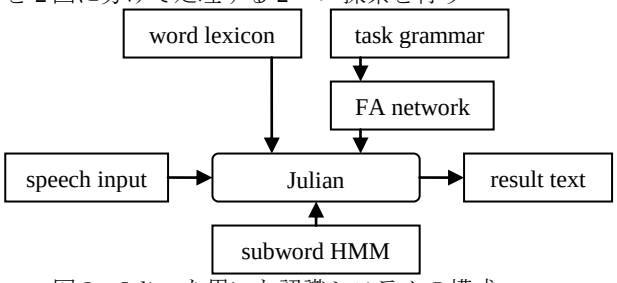

図 2　Julian を用いた認識システムの構成

3.1.1. タスク作成

音楽再生機器における曲名やアーティスト名検索を想定し，アーティスト名･曲名などを発話した音声を認識するタスクを作成した．このタスクでは，「○○○(アーティスト名)」「■■■(曲名)」といった文章を認識できる．

3.1.2. 単語辞書

認識可能な単語数として，アーティスト名･曲名を合わせ，約 5000 単語を登録した．この単語数をさらに増やした場合，認識結果として出力される候補が増えるため，認識率に変化が現れることが考えられる．そのため本研究では，単語数を多くした場合の認識率も求め，認識率がどのように変動するのか，またどの程度の単語数がこのシステムに適しているのかなども考察した．

4. 集音機器の性能評価

音声による機器操作に関連する文献[1,2,3,4,5]には，集音機器に関する詳しい記述はない．そのため，まず音声の集音に使用するマイクに着目した．機器操作を行うための音声命令を集音する機器としては，ダイナミックマイクのようなタイプよりは，より小型のマイクの方が実用面において適切であると考えられる．あるいは，2 章で述べた課題の面から，比較的雑音を拾いにくいマイクである必要もある．本研究ではこれらを考慮し，小型かつ比較的雑音を拾いにくいヘッドセット型マイクを集音機器とした．このマイクは集音時に話者の右耳に装着するものであり，長所としては機器操作を行う上で比較的手間がかからず使いやすい点が挙げられる．

本章では，このヘッドセット型マイクで収録した音声を認識させ，その認識率を比較的フラットな特性を持つダイナミックマイクで収録した場合の認識率と比較することで，ヘッドセット型マイクの集音機器としての性能評価を行った．音声の収録は雑音の少ない静かな教室を使用して行ったが，本研究の目標とするシステムは雑音の多い環境で利用されることが想定されるため，雑音の多い場所でも音声を収録し，認識率を出して考察を行った．この結果から，ヘッドセット型マイクが，機器操作を行う上で適切な集音機器であるかどうかという判断ができる．

4.1. 集音機器

比較に使用する 2 種類のマイクを図 3，図 4 に示す．ダイナミックマイクとしては audio-technica DYNAMIC VOCAL MICROPHONE VD3 を，ヘッドセットマイクとしては Sony Ericsson Bluetooth Headset HBH-PV705 を使用した．尚 HBH-PV705 は，AVRCP，HSP，HFP，DUN，A2DP 等のプロファイルを持っている．

図 3　VD3

図 4　HBH-PV705

4.2. 評価方法

1 つの音声をヘッドセットとダイナミック両方のマイクで同時に収録する．その後収録した音声を Julian で認識させ，それぞれの認識率を求める．この認識率の違いを比較することによってマイクの性能を評価する．また，ヘッドセットマイクを用い雑音の多い実環境下でも収録し，認識率がどの程度であるかを考察する．

4.3. 評価用音声の収録環境と収録方法

まず雑音の無い環境で収録を行うために，誰もいない教室を使用して収録を行った．

被験者にヘッドセットマイクを装着してもらい，同時にダイナミックマイクをマイクスタンドにセットする．収録開始の合図とともにダイナミックマイクに向けて発話してもらい，1 つの音声をこれら 2 つの収録機器で同時に収録する．尚，ダイナミックマイクはサンプリング周波数 16kHz，ヘッドセットマイクはサンプリング周波数 8kHz での収録である．

さらに，実際にこのシステムを利用する場面に近い環境でも音声を収録するために，以下の 4 種類の環境で収録を行った．

1. 車内 1(走行中)　　2. 車内 2(走行中，音楽あり)
3. 路上　　4. TV の前

1 は走行中の自動車の車内である．エンジン音や，他の自動車の走行する音などの雑音がある．2 は車内に音楽を流した状況での収録である．3 は環状七号線沿いの

コンビニエンスストアの駐車場である．自動車やバイクが走行する音や風などの雑音がある．4 は電源の入っているテレビの前での収録である．テレビの音声が主な雑音である．

尚，この場合は実環境下でのヘッドセットマイクの認識率のみを求めることが目的であるので，ダイナミックマイクでの収録は行っていない．

4.4. 評価用音声

発話内容については，文献[6]において使用されていた音声サンプル 50 個のうち 25 個を選択し，それらと同じ内容とした．これらは「ケレル」「クレージーケンバンド」等，全てアーティスト名，あるいは曲名のみを発話したものである．

雑音の無い環境では，1 人につき 25 単語をそれぞれ 10 発話ずつ，つまり 1 人につき計 250 発話を収録した．これをのべ 10 人分収録し，1 単語につき 100 発話の計 2500 発話を認識用サンプルとした．また実環境下では，1 人につき 25 単語を 2 発話ずつ，つまり 1 人につき計 50 発話を収録した．話者の人数は 1，2，3 が 3 人，4 が 1 人であり，1，2，3 については計 150 発話ずつ，4 については計 50 発話を認識用サンプルとした．

4.5. 認識結果と評価

認識率は，話者によって認識率に違いがあるのか，話者によらず認識率の低い単語があるのかなどを併せて調べるため，単語別，話者別にまとめた．単語によってはダイナミックマイクの認識率の方が高かった単語もあったが，全体的な認識率を比べた結果としては，ヘッドセットマイクで収録した音声が約 75%，一方ダイナミックマイクの方は約 73%となり，ヘッドセットマイクの認識率の方が高いことが確認できた．ヘッドセットマイクがダイナミックマイクと比べて同等かそれ以上の性能を持つことが確認でき，小型であることや使いやすさを考慮すると，このマイクが機器操作を行うための集音機器として不適切でないことが言える．

尚，実環境下で収録した音声は，5 章で後述する単語辞書修正後に認識を行い，認識率を求める．

5. 認識率の向上

4 章においてヘッドセットマイクで収録した音声の認識率をさらに向上させるため，比較的正しく認識されにくい音素を含む単語について，単語辞書の修正を行った．

5.1. 発音の揺れ

認識率を単語別に見た場合，比較的認識率の低い単語には長音や”w e”，”w i”を含む場合が多く見受けられた．長音を含む単語の音素列は，a : , i :, などと” : ”を用いて表されるが，これだけでは不十分と考え，それぞれ a a, i i のように音素を重ねた場合の音素列も追加登録をした．また”w e”，”w i”を含む単語に関しては，”w”を”u”を書き換えたものを追加登録した．表 1 参照．

表 1　長音や”w e”，”w i”を含む単語の音素列例

単語	音素列
ビートシティー	b i: t o sh i t i:
	b i i t o sh i t i i
ジェフウェイン	j e f u w e i N
	j e f u u e i N

5.2. 無声音で始まる単語

[s]などの無声子音が単語の冒頭に来た場合，それに続く狭母音(特に u，i など)は無声音になる傾向がある．このことから，[s]で始まる単語は 1 文字目の”s”が分析できない場合があると考えた．実際に無声音で始まる単語”スマイルアゲイン”を選び，それぞれのスペクトログラムを比較した．図 5 参照．

左図は文献[6]で使用されていた音声サンプルであり，サンプリング周波数 16kHz で収録されたものである．また，右図は本研究内で収録・使用している音声サンプルであり，サンプリング周波数 8kHz で収録したものである．楕円で囲んだ部分が無声音の周波数を示した部分であり，見比べると右図では最初の無声音が表示されていないことがわかる．つまり，8kHz で収録を行った場合は，冒頭の”s”は分析されにくいようである．このため無声音で始まる単語については，無声音の部分を削除した場合の音素列を追加登録した．表 2 参照．

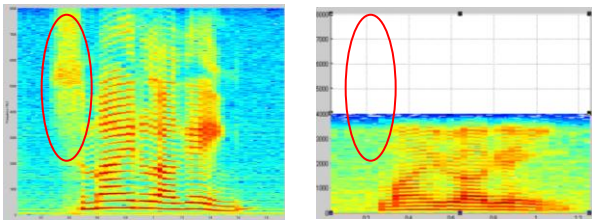

図 5　16kHz で収録した音声と 8kHz で収録した音声のスペクトログラムの違い

表 2　無声音で始まる単語の音素列例

単語	音素列
スマイルアゲイン	s u m a i r u a g e i N
	m a i r u a g e i N
ストレートアヘッド	s u t o r e: t o a h e q d o
	s u t o r e e t o a h e q d o
	t o r e: t o a h e q d o
	t o r e e t o a h e q d o

5.3. 認識率

本章では，例として挙げた単語を含め，発音の揺れに関しては 16 単語，無声音に関しては 4 単語において同様の修正を加えた．その結果，これらの単語の認識率は向上し，全体の認識率が約 85%へ向上した．今後もこの認識率を向上させていく上で，この修正は非常に有効な手段であると言える．例として挙げた単語の認識率の変化を表 3 に示す．

表 3　辞書修正を加えた単語の認識率の変化例

単語	修正前	修正後
ビートシティー	20%	83%
ジェフウェイン	20%	83%
スマイルアゲイン	59%	62%
ストレートアヘッド	54%	81%

6. 実環境下で収録した音声の認識率

単語辞書の修正後，4 章で述べた実環境下で収録した音声について認識を行い，環境毎の認識率とその SNR を求めた．表 8 参照．SNR の計算方法は以下の式により，音声の最初の 100 ミリ秒に含まれる雑音のパワーの平均

を P_{noise}，音声の発話部分のパワーの平均を P_{signal} として計算した．

$$SNR[dB] = 10\log_{10}\left(\frac{P_{signal}}{P_{noise}}\right)$$

P：パワー

雑音の無い場所で収録した音声の認識率と比較すると，やはり雑音の多い実環境下での認識率の方が，全体の認識率は低かった．尚，雑音の無い環境で収録した音声の SNR は 16.32 であり，これを考えると，認識率はある程度 SNR の大きさに依存していると言える．

表 4　実環境下での認識率と SNR

	車内 1	車内 2	路上	TV
認識率	77%	79%	71%	88%
SNR	11.88	13.93	11.00	14.85

7. 単語辞書の総単語数の増加

5 章で出した結果は，5000 単語の辞書を使用して認識率を出したものである．この単語数をさらに増やした場合，認識結果として出力される候補が増えるため，認識率に変化が現れることが考えられる．この変化を考察するために，単語辞書の単語数を 10000 単語，20000 単語に増やした場合の認識率を求めた．それぞれの認識率の違いを表 5 に示す．この結果，単語総数の増加とともに認識率が低下することがわかる．特に路上では認識率が 6 割程度に低下し，話者別にみると認識率が 5 割に満たない話者も出た．

表 5　単語辞書の単語数と認識率

単語数	雑音無し	車内 1	車内 2	路上	TV
5000	84.4%	77%	79%	71%	88%
10000	80.4%	73%	75%	64%	82%
20000	77.7%	67%	73%	61%	82%

8. 正解の順位

認識候補を 1000 位まで表示させることで，誤認識した単語では正解が認識候補の何位に表示されているのか，1 位にはどのような単語が表示されているのかなどを調べた．認識率の低い単語は，やはり正解が順位の低いところに出てくる傾向にあったが，認識率は低いが毎回上位に正解が現れる単語もあった．たとえば，特に認識率の低かった”シングアポケットビスケッツ”と”スマイルアゲイン”について，前者では誤認識される場合のスコアは大抵低く，正解の順位はかなり低いものが多かった．しかし後者ではスコアは大抵高く，正解の順位が極端に低いことはほとんどなかった．また実環境下で収録した音声については，雑音の大きいサンプルは誤認識しやすく，スコアの値は低く正解の順位もかなり低い傾向にあった．

単語辞書の単語数を増やした場合は，5000 単語の辞書使用時には 1 位に正しい認識結果が出力されたサンプルが 10000 単語の辞書に変更後に 1000 位以下に出力されるサンプルもあった．これは 10000 単語から 20000 単語に変更した場合にも見られ，探索パラメータの変更により正しい認識候補が途中で枝切りされたためと考えられる．

9. 考察

ヘッドセットマイク収録した音声の認識率は，ダイナミックマイクで収録した音声の認識率よりも高いことを確認した．これは収録の際，ヘッドセットマイクの方が話者の口に近い場所にセットされていることが理由として考えられる．この結果から，ヘッドセット型マイクはダイナミックマイクよりもこのシステムに適していることが言える．また，認識率向上には単語辞書修正は有効な手段であることがわかった．認識しやすい音素列，認識しにくい音素列を理解し，認識しにくいものについては複数の音素列を登録して対応することが，認識率向上において重要である．

単語辞書の登録単語数については，単語数の増加に伴い認識率が下がることがわかる．学生 10 名における音楽再生機器の登録曲目数等を調査した結果，登録数はアーティスト数，曲数，アルバム数を合わせ，最大でも 5000 を超えないことが分かった．この結果から単語辞書の単語数は 5000 単語あれば十分であると考えられ，単語数を 10000 単語，20000 単語に増やした場合の認識率低下を考慮すると，単語数を 10000 単語やそれ以上に増やす必要性は無いと言える．

また，ヘッドセットマイクを使用し実環境下で収録した音声が，7 割から 8 割程度の認識率を出せることを実験により確かめた．このことからもヘッドセットマイクの性能がある程度確認でき，集音機器としての機能を十分に果たせることが言える結果となった．

10. あとがき

本研究では，ヘッドセットマイクで収録した音声の認識率を求めることで，ヘッドセットマイクの集音機器としての性能評価を行い，このシステムに適していることを示した．雑音のある環境での認識率向上，認識結果を受けて AV 機器を操作するシステムの構築などが今後の課題となる．

文　献

[1] 武藤他，”PC を介したロボット･情報機器の音声操作に関する検討”，信学技報．HCS，ヒューマンコミュニケーション基礎，Vol.100, No.712(20010314) pp. 1-6, HCS2000-56

[2] 長田他，”自然言語を用いて家庭機器操作を行う対話システム”，信学技報．SP，音声，Vol.98，No.317(19981015) pp. 23-30

[3] 河原崎他，”音声認識による家電機器のリモコン制御(福祉工学･機器 1)”，福祉工学シンポジウム講演論文集，Vol.2004(20040912) pp. 197-200

[4] 河原崎他，”音声による家電機器のリモコン操作”，バイオメカニズム学術講演会予稿集，Vol.25 (2004/10/23-24) pp. 215-218

[5] 福島他，”視覚障害者のための家電機器操作用マルチリモコンの開発”，埼玉県産業技術総合センター研究報告，Vol.4 (2005 年度) pp. 23-27

[6] 重田武弘，”項目反応理論を用いた大語彙単語認識の期待正解率の推定”，07 年度法政大学情報科学研究科修士論文．