TR-IT-0355

# 雑音抽出に向けてのＨＴＫに於ける性能評価

池田 陽平　　北川 敏　　Nick Campbell

2000.2

## 概要

CHATR のデータベース作成のために、録音した音声を人手でラベリングしていくには膨大な時間と労力がかかり、個人によって差も出てくるため、自動化することが望ましい。そこで本報告では、音声認識ツールであるＨＴＫでの自動ラベリングを目指せるようその下準備としてラベリング機能の評価を行った。

# もくじ

# 第 1 章

# はじめに

現在、ATR 音声翻訳通信研究所では、CHATR のデータベース作成のため録音した音声から雑音の部分を取り除き、音素ごとにラベル付けを行う作業を行っている。

録音の内容は、５秒程度の文が録音されており、テキストを読んでいる声のほか、文と文の間にテキストのページをめくる音や咳、読み間違えなどの雑音が入っている。この録音を雑音が入らないよう１文ずつに切り分け、音素ごとにラベル付けする作業には膨大な時間と労力がかかる。また、ラベラーによってある程度の個人差がでてしまう。

よって、この作業を音声認識ツールである HTK を用いて実現することが目標である。

そこで、本報告書では、ＨＴＫでの自動ラベリングを目指せるようその下準備としてラベリング機能の評価を行った。

# 第 2 章

# 音素単位でのラベリング

最終的な目標はCHATRのデータベースを作ること。つまりテキストと録音テープからできるだけ正確にラベルしていくことである。まずは音素をできるだけ正確な場所にラベリングできないと、目標である自動ラベリングが達成できない。そこで、HTKがどの程度音素をラベルできるか試験した。

## 2.1 試験方法

HTKに正解の基準となるラベラーがつけたラベルを学習させ、ATRの基準による音素単位でのラベリング性能を試験した。

試験はオープンとクローズの2種類とし、用いる音声はMYSとFUMの男女各一名とした。また、MYSとFUMを比較できるように同じ内容を話している文で試験をした。

1. ＭＹＳ（男性）

試験に使った音声とそれに含まれる音素数

オープン試験の学習には

- MYS_503_A_10 から MYS_503_A_50
- MYS_503_B_01 から MYS_503_B_50、音素数５７６８

を使用し、認識には

- MYS_503_A_01 から MYS_503_A_09、音素数４４３

を使用した。

また、クローズ試験の学習には

- MYS_503_A_01 から MYS_503_A_50
- MYS_503_B_01 から MYS_503_B_50、音素数５７６８

を使用し、認識には

- MYS_503_A_01 から MYS_503_A_09、音素数４４３

を使用した。

2. ＦＵＭ（女性）

試験に使った音声

オープン試験の学習には

- FUM_503_001 から FUM_503_099、音素数５３２５

を使用し、認識には

- MYS_503_001 から MYS_503_009、音素数４４３

を使用した。

また、クローズ試験の学習には

- FUM_503_001 から FUM_503_099、音素数５７６８

を使用し、認識には

- FUM_503_001 から FUM_503_009、音素数４４３

を使用し、ALIGN した。

## 2.2 結果及び考察

試験結果を次に示す。出した値は次の通りである。

- 正解より 0.05 秒以上ずれている割合
- 正解より 0.10 秒以上ずれている割合
- 平均で 0.02 秒以上ずれている音素

  各音素ごとにずれを合計し平均したもの

表 2.1: ＭＹＳの結果

| | クローズ | オープン |
|---|---|---|
| 正解より 0.05 秒以上ずれる | 9.03% | 9.48% |
| 正解より 0.10 秒以上ずれる | 7.22% | 7.67% |
| 平均で 0.02 秒以上ずれる | U,f,j,k,s,sh,w,y | U,f,j,k,s,sh,w,y |

表 2.2: ＦＵＭの結果

| | クローズ | オープン |
|---|---|---|
| 正解より 0.05 秒以上ずれる | 6.32% | 6.55% |
| 正解より 0.10 秒以上ずれる | 4.06% | 4.06% |
| 平均で 0.02 秒以上ずれる | U,w,y | U,kk,w,y |

結果は、図２、１に示すように音素がずれているところを探すのが難しいほどよい結果が出たといえる。この結果より、文章にもよるが 100 文、5000 音素程度の学習で音素のラベリングはできることがわかった。

平均で 0.02 秒以上ずれる音をあげてあるが、実際 CHATR で使用するなら 0.02 秒ぐらいならそんなに問題はないと思います、0.05 秒以上ずれる音にすると'w' がＭＹＳ、ＦＵＭ両者にあがる。（付録参照）

また、学習させたとき bb,dd,ff,gg,hh,jj,mm,nn,rr,tts,v,vv,ww,yy,zz の各音素が学習するには出現回数が少なすぎる（HInit Error ; Too Few Observation）とでたが、これらの単語は母音などより使われ方が少なく、少ない学習でも学習できると考えた。

ＭＹＳの ALIGN 結果の例を次にしめす。これは１番目の文「あらゆる現実をすべて自分の方へねじまげたのだ。」である。一番上は音声信号、上から２番目はオープンの ALIGN 結果、下から２番目はクローズの ALIGN 結果、一番下がラベラーが付けたラベル（正解）です。全体的なラベルの振り方を見てください。また、次には拡大したものを載せています。ＦＵＭの１番目の文の ALIGN 結果の例も付録に付けてあります。

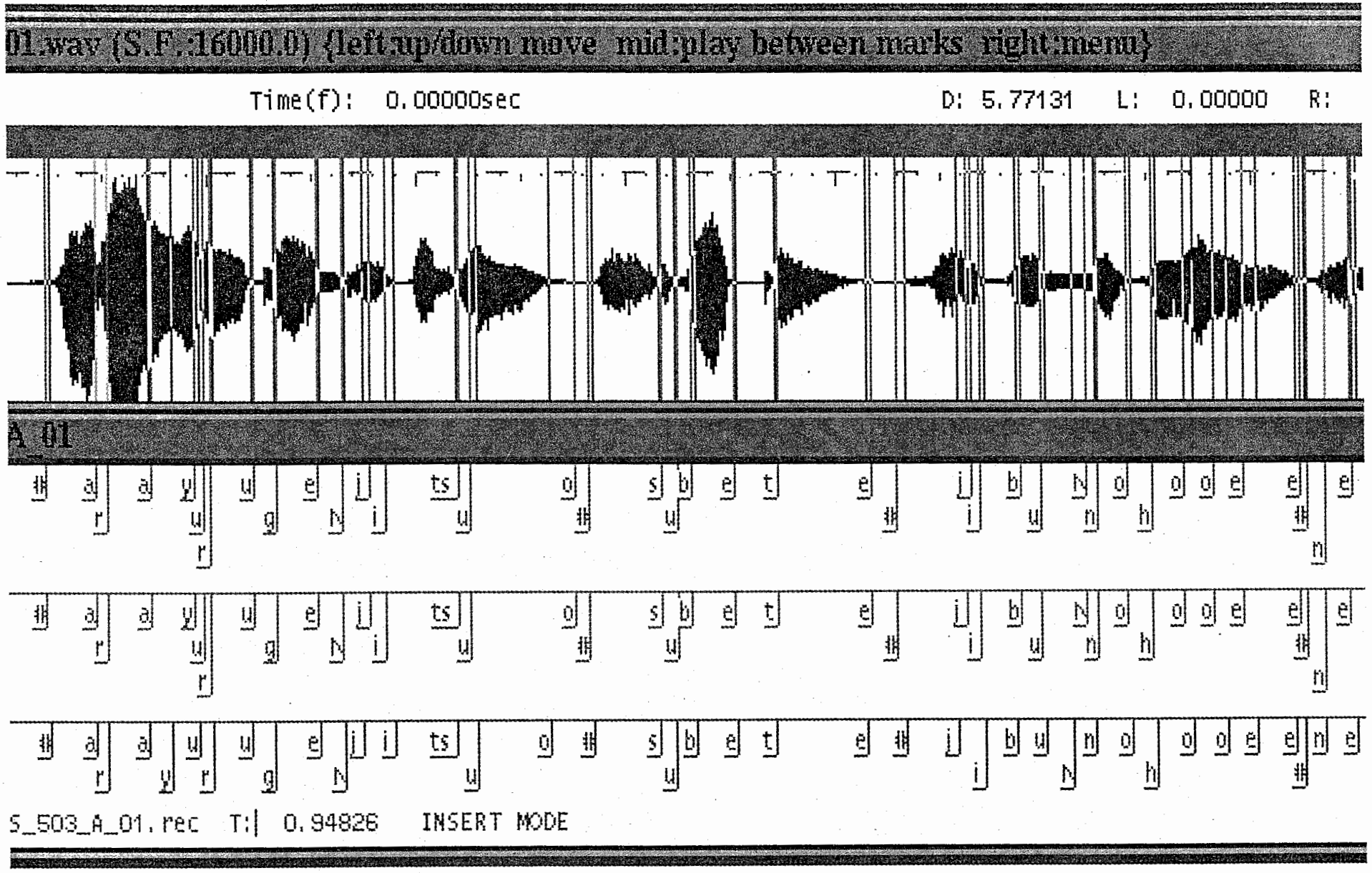



図 2.1: MYS の ALIGN 結果

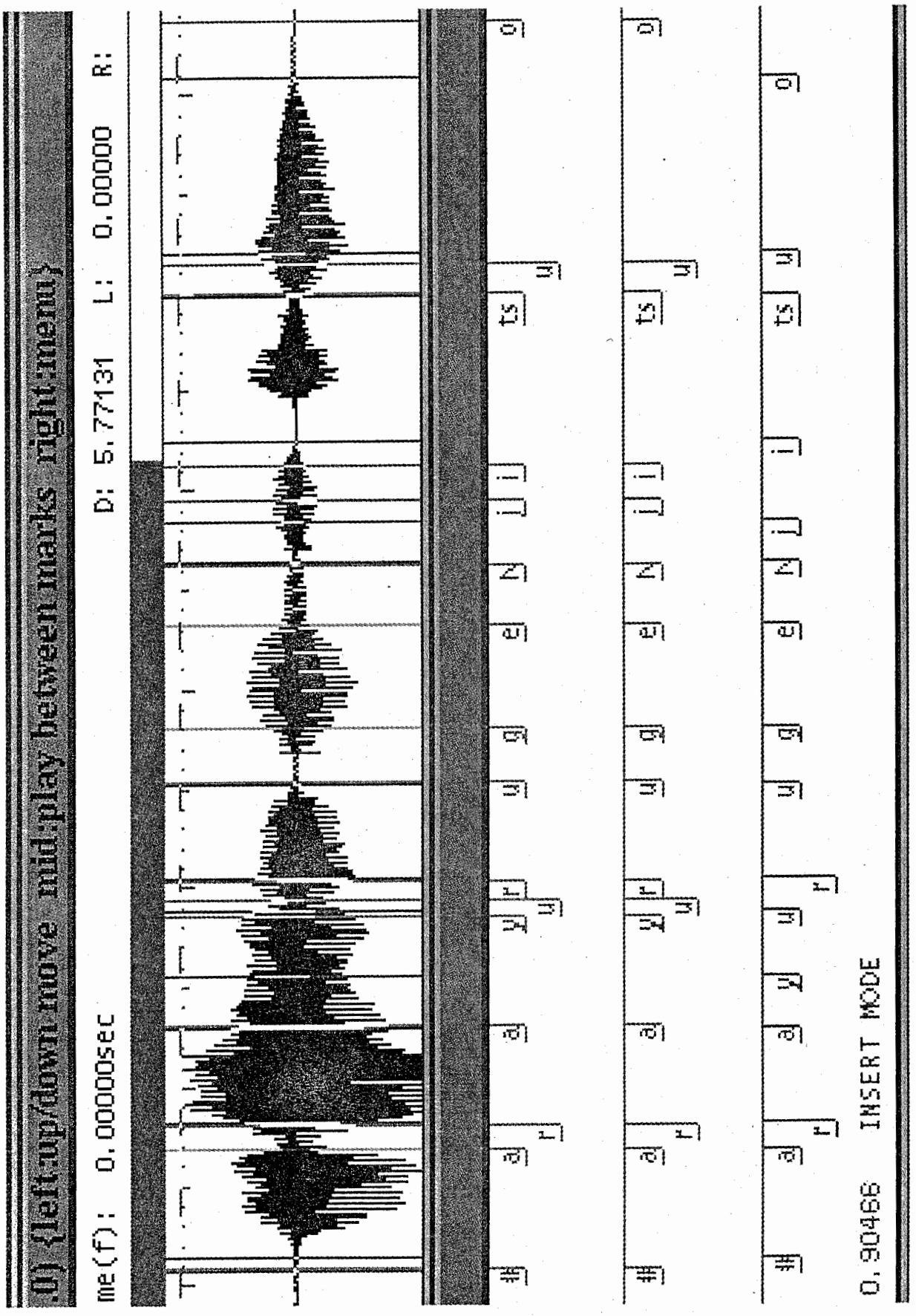



図 2.2: MYSのALIGN結果、拡大版

# 第 3 章

# 辞書を使ってのラベリング

## 3.1 辞書とは

次は辞書を用いて認識させました。

そこでまず、辞書とはどんなものかであるが、基本的には１つの文を品詞ごとに分けたり、アクセンチュアルフレーズごとに分け、それを１つのファイルにまとめたものです。

例　　老人ホームの場合は、健康器具や膝掛けだ。

$\Longrightarrow$　　老人ホーム　の　場合　は　健康器具　や　膝掛け　だ

辞書に入っている「文を分けられたもの」を「単語」と呼ぶことにします。

次に辞書を使っての認識ですが、２章のような音素単位ではなく、辞書に入っている単語単位で最も対数確率（対数尤度）の高いものを認識結果として出力します。

今回は

/DB/CHATR01/chatr_dbs/MYS503q_tobi-2/others/MYS.words

を基準として分けていった。

## 3.2 試験方法

試験に使ったデータと試験方法は音素での実験のときとまったく同じである。

つぎに、文の区切り方にはいろいろな考え方があると思いますが、今回は４ ～ ６単位に分けました。なぜ４ ～ ６単位なのかは４章で述べます。とりあえず今は、試験に使った文は４ ～ ６単位で分けると、自立語＋付属語もしくは自立語のみというようにきれいに分れるからとしておきましょう。

## 3.3 結果及び考察

試験結果を次に示す。出した値は単語単位での認識率である。

認識に使われる文の単語総数は４４語である。

辞書内に入っている単語数は２９００語。これはMYS_503_のすべての文から作成したものである。

表 3.1: ＭＹＳの結果

| | クローズ | オープン |
|---|---|---|
| MYS | 81.82% | 77.27% |
| FUM | 75.00% | 63.64% |

音素の認識率より極端に下がりましたが、理由が2点ほど考えられます。

一つは辞書のサイズが２９００語と大きいため。認識で使われる単語数４４語ですればもっと上がるはず。

もう一つは学習させる文章数が少ないこと。１００文で約５００語ですが、２０００語は必要とのことです。

辞書を使った認識で正解の結果が出た例と間違いが出た例を載せておく。

正解が出た例は図3.1で、話者FUMで8番の文をオープンで認識したものです。文の内容は「老人ホームの場合は、健康器具や膝掛けだ。」です。1段目がピッチ、2段目が辞書での認識結果、3段目がオープンで音素での認識結果、4段目がラベラーが付けたラベル（正解）です。

間違いが出た例は図3.2で、話者MYSで5番の文をオープンで認識したものです。文の内容は「救急車がじゅうぶんに動けず、救助作業が遅れている。」で、拡大して後半を載せています。1段目がピッチ、2段目が辞書での認識結果、3段目がオープンで音素での認識結果、4段目がラベラーが付けたラベル（正解）です。

「kyuujo」のところが「fukuya」と間違って認識されています。これは単に「kyuujo」のほうが「fukuya」より尤度が高かったと考える以外にはないのですが、「kyuujo」のほうの尤度もある程度は高いと思います。

今後の課題でもあるのですが、まだ尤度（確率）を出せていません。4章では、確率を使って雑音検出をする案を出しています。

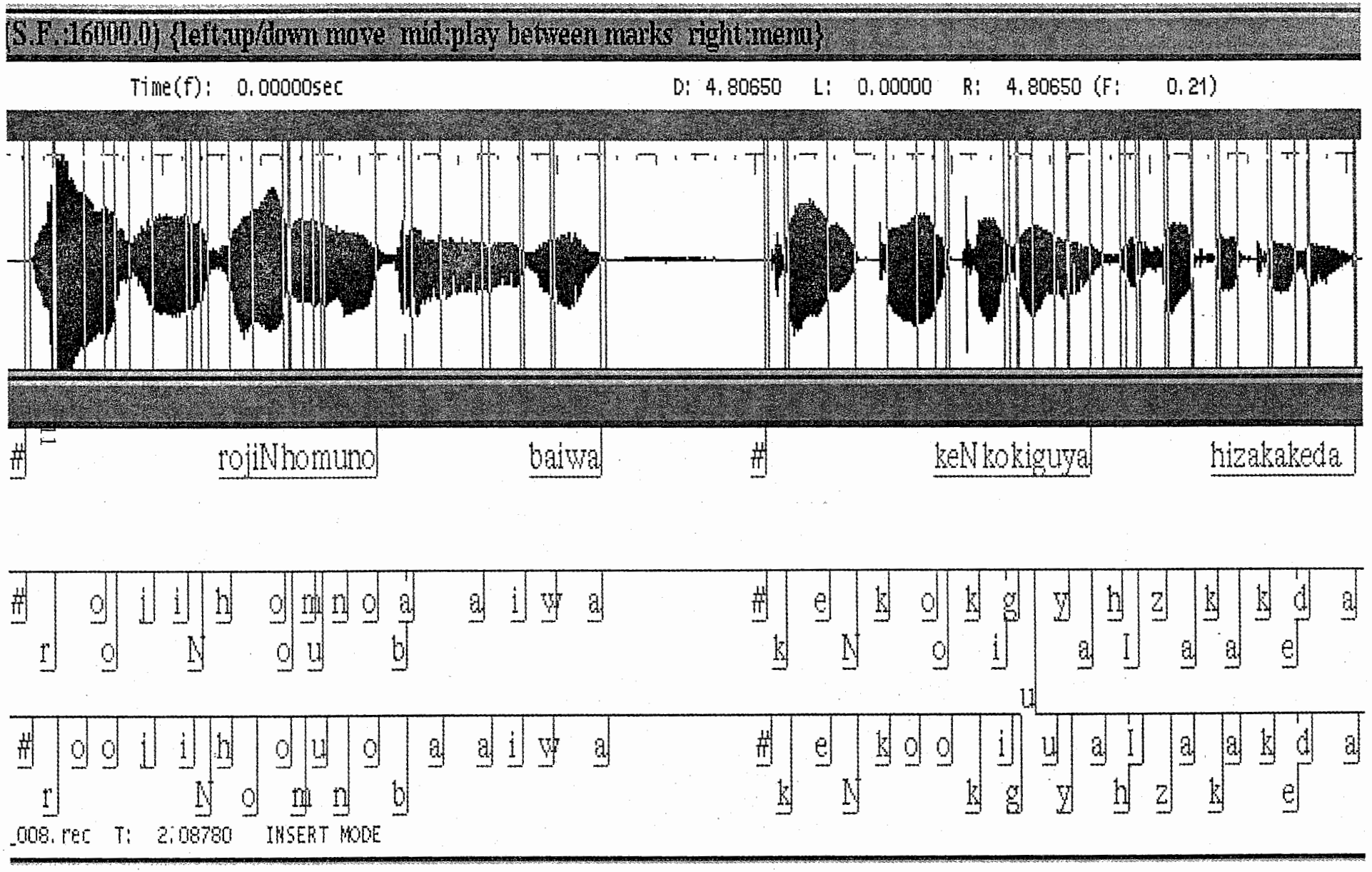

S.F.:16000.0} {left:up/down move mid:play between marks right:menu}
Time(f): 0.00000sec
D: 4.80650 L: 0.00000 R: 4.80650 (F: 0.21)
# rojiNhomuno baiwa # keNkokiguya hizakakeda
_008.rec T: 2:08780 INSERT MODE

図 3.1: 正解例

0 {left:up/down move mid:play between marks right(menu)}
e(f): 0.000000sec D: 5.55350 L: 0.00000 R: 5.55350 (F: 0.18)
fukuya
sayoga
okureteiru
.11574 INSERT MODE

図 3.2: ミスの例

# 第 4 章

# まとめ

## 4.1 まとめ

- HTK の音素を ALGIN する機能が CHATR で使える程度はあるということ。
- 辞書認識については、CHATR のデータベースのために、信頼性の高い語を選びだせそうなことがわかった。

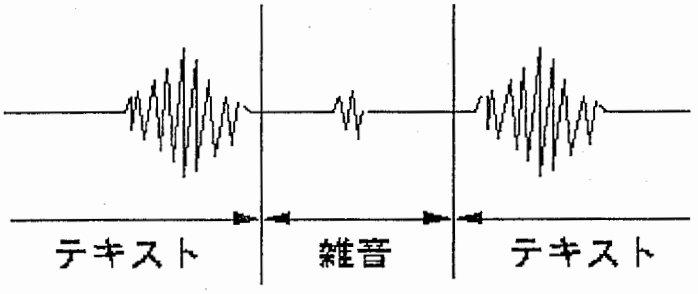
## 4.2 今後の課題

今後の課題としてアイデアを残しておきます。

このアイデアは HTK の出力で辞書認識時の確率が出せることが前提です。

まず、図 4.1 を見てください。

なぜ辞書での認識が必要か？




ＨＴＫで認識時の確率を出せるなら

雑音は　　音素で認識・・・認識される確率が高いかもしれない

　　辞書で認識・・・認識される確率が低いと考えられる

図 4.1:

テキストの間に雑音が入っていると考えてください。HTK でこの文を音素単位で認識したとき、２章の結果からもわかるようにかなりの確率でテキスト部も雑音部も認識されるでしょう。雑音部も音素単位ならマッチするかもしれないからです。

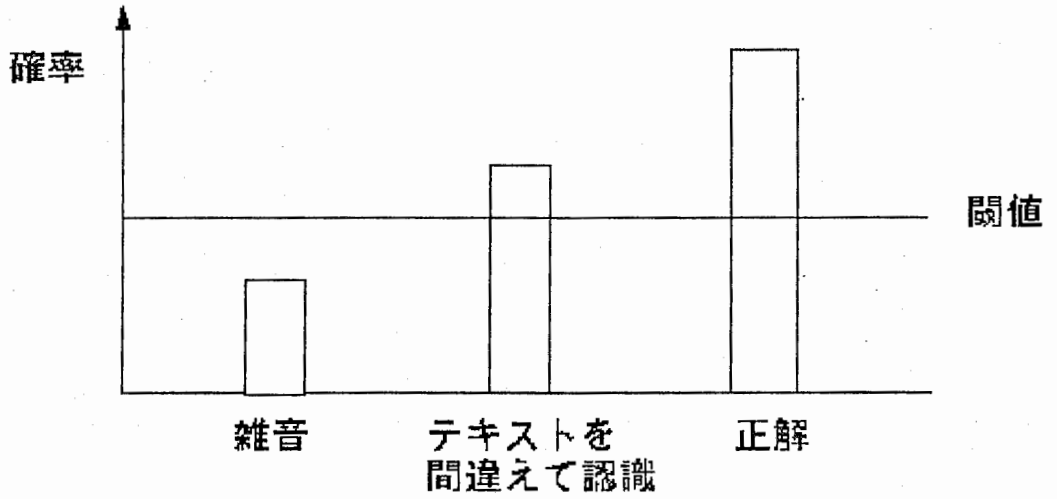
次に辞書を使って認識したとき、どうなるでしょうか。辞書には雑音が入っていないので、雑音部は無理やりいちばん近いと計算される単語が持ってこられ、確率（尤度）はかなり小さいはずです。３章の結果よりテキスト部でも間違える可能性はありますが、間違えたところは、本当の単語を越える確率を持ってたわけですから、雑音部よりは確率は大きいはずです。

つまり認識時の確率の大きさを比べると、

単語を正解したときの確率　＞　単語を認識ミスしたときの確率　＞　雑音に無理やり単語をつけた確率

となるはずです。あとは図 4.2 のような閾値さえ設定できれば雑音混じりの部分は検出できるのではないかと考えました。少しぐらい閾値を高めにとってでも、信頼性の高いデータが手にはいればそれでいいと考えます。

辞書を使えば




となる関係があるかも、、、

図 4.2:

次に文の区切り方です。

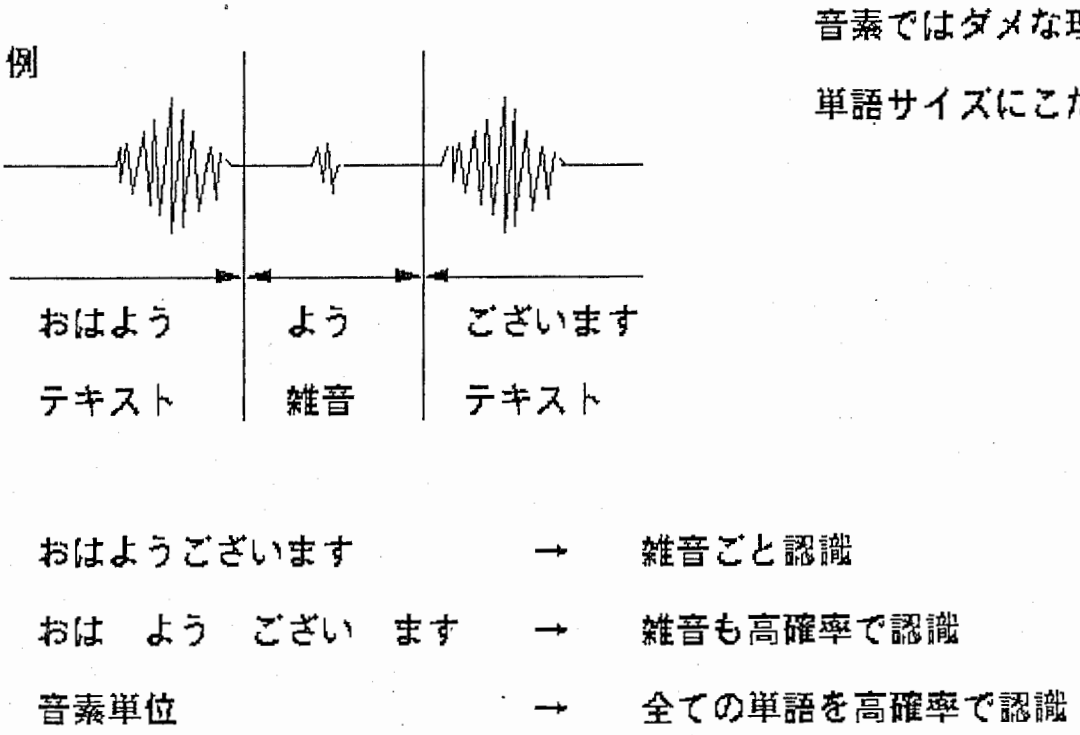



図 4.3:

例は「おはようございます」と録音しているときに、たまたま向うから友達が来て「よう」と言われた、という設定です。このとき、いちばん理想的な単語の分け方は、

おはよう　ございます

となります。「おはようございます」を１つの単語にすると、雑音ごと認識してしまう可能性があります。「おは　よう　ござい　ます」と４つの単語に分けると、辞書に「よう」という単語が入ってしまい、雑音部も高確率な認識率を残してしまいます。音素単位は全ての音素を正しく認識するという結果が出ているので、使えません。今回の実験では単語が長くもなく、短くもないように切ると、４～６単位がちょうどよい値になると考えました。

# 謝辞

本研究を進めるにあたって、暖かく見守りつつ、多くの御指導を頂いた Nick Campbell 第二研究室室長に心から感謝致します。

さらに、本研究に対する適切な助言を頂き、様々な相談にのって下さった ATR 音声翻訳研究所第二研究室の皆様、深く感謝致します。

1999 年 9 月 24 日

池田 陽平

# 付録 A

# 各データの所在（出現順に）

1章で使用したデータ

MYS のデータ

/DB/CHATR01/chatr_dps/MYS/

FUM のデータ

/DB/CHATR04/chatr_dps/FUM/

音素を ALIGN した結果（MYS、オープン）

/home\as59\xyikeda\MYS_HTK\op_M_ph9_ALIN/

音素を ALIGN した結果（MYS、クローズ）

/home/as59/xyikeda/MYS_HTK/clo_M_ph_ALIN/

音素を ALIGN した結果（FUM、オープン）

/home/as59/xyikeda/HTK/op_F_9_ALIN/

音素を ALIGN した結果（FUM、クローズ）

/home/as59/xyikeda/HTK/F_Mon_ALIN/

ALIGN した音素のずれを計算するプログラム

/home/as59/xyikeda/HTK/propererror.pl

各音素の ALIGN したあとの正解からのずれ（MYS、オープン）

/home/as59/xyikeda/MYS_HTK/op_M_ph9/_ALIN/newrec/length/diff/average

各音素の ALIGN したあとの正解からのずれ（MYS、クローズ）

/home/as59/xyikeda/MYS_HTK/clo_M_ph_ALIN/newrec/length/diff/average

各音素の ALIGN したあとの正解からのずれ（FUM、オープン）

/home/as59/xyikeda/HTK/op_F_9_ALIN/newrec/length/diff/average

各音素の ALIGN したあとの正解からのずれ（FUM、クローズ）

/home/as59/xyikeda/HTK/F_Mon_ALIN/newrec/length/diff/close_ave

2章で使用したデータ

使用した辞書

/home/as59/xyikeda/HTK/dict/MYSsec6

使用したネットワーク

/home/as59/xyikeda/HTK/dicts/netMYSsec6

辞書を使っての認識結果（MYS、オープン）

/home/as59/xyikeda/MYS_HTK/op_M_ph9/_ALIN/

辞書を使っての認識結果（MYS、クローズ）

/home/as59/xyikeda/MYS_HTK/clo_M_ph_ALIN/

辞書を使っての認識結果（FUM、オープン）

/home/as59/xyikeda/HTK/op_F_9_ALIN/

辞書を使っての認識結果（FUM、クローズ）

/home/as59/xyikeda/HTK/F_Mon_ALIN/