音声言語情報処理 15-21
(1997. 2. 8)

音声波形からの音素片記号系列を用いた音声要約と話題要約の検討

中沢正幸　遠藤隆　古川清　豊浦潤　岡隆一

新情報処理開発機構　つくば研究センタ
〒305　茨城県つくば市竹園1-6-1つくば三井ビル14F
TEL 0298-53-1676　E-mail nakazawa@trc.rwcp.or.jp

あらまし　筆者らは、音声中の互いに類似した十分な長さの区間という音声表層の情報から、要約の生成を通して意味の深層情報の情報を得るという新しいアプローチを試みている。このアプローチは、音声全体や文全体を理解をするという深い意味や内容理解には立ち入らずに、音声中に現われる単語の表層情報を手がかりにシソーラス等の言語情報を用いて話題の推定を行うという、いわば斜め聞きというべきものである。本稿では、任意話題の音声・話題要約を実現する過程で、音声波形の表層情報と意味の深層情報を結びつけるために必要不可欠な音素片記号系列を対象とした20万語規模の大語彙単語認識を高速に行う手法を提案する。
キーワード　音声要約、話題要約、音声認識、音素片、重要語抽出、単語認識、大語彙認識

A STUDY ON SPEECH SUMMARY USING DEMI-PHONEME SYMBOLS GENERATED FROM SPEECH WAVES

Masayuki Nakazawa　Takashi Endo　Kiyoshi Furukawa　Jun Toyoura　Ryuichi Oka

Tsukuba Research Center, Real World Computing Partnership
Tsukuba Mitsui Building 14F, 1-6-1 Takezono Tsukuba-shi, Ibaraki 305
TEL 0298-53-1676　E-mail nakazawa@trc.rwcp.or.jp

Abstruct　We propose a new approach to the speech interface based on superficial speech information and natural language information through generating a topic summary. We know, it is difficult for system to understand a spontaneous speech. The major problem of speech recognition is that difficulty of creating grammar and using as a man-machine interface. This paper describes a method for segmentation of topics from a spontaneous speech using by superficial speech information and natural language information, and how to recognize for large vocabulary using demi-phoneme symbol series.
Keywords　speech summary, topic summary, speech recognition, demi-phoneme symbols, word recognition, large vocabulary

1. はじめに

近年多数のデータベース、特に定期的に刊行される雑誌記事などの全文データベースが一般ユーザに対して解放されるようになった。これらに効率よくアクセスするためのオンライン情報提供サービスも一般化しつつある。インターネットにおけるサーチエンジンなどはその典型的な例だと言える。

一方以前から、DAT・MD・カセットテープなどは録音メディアの普及により、ラジオ・テレビ放送、会議などの人間同士の対話である音声情報を簡単にかつ大量に保存することが可能になってから久しい。しかし、この音声情報に対しては情報提供サービスが一般化していないばかりか、効率よくアクセスするための検索手段も用意されていない。このことは、ユーザが大量の情報に直面した時に、情報検索の負担の増加という問題が生じることを指している。

この時重要なのは、ユーザが必要としているのは、直面した情報の詳細な内容ではなく、その情報のあらすじで十分だということである。それだけで、情報の範囲を狭めることができ、効率的な検索ができるようになるからである。

筆者は、人間同士の自然対話音声を対象として、その対話の音声・話題要約に関する研究を行っている。自然対話を要約することにより、その要約をもとに、ユーザに情報検索の手がかりを与えようという試みである。すなわち、音声表層情報から抽出した結果である要約を手がかりに、ユーザが情報スキャニングを行い、希望する情報のみを簡単に情報の洪水の中から拾い出すことを可能にしようというわけである。

このアプローチを実現するために、話題中の重要な単語を音声中の互いに類似した十分な長さの区間と仮定し、その区間を抽出することで、話題を反映した重要単語を抽出する方法を提案した[中沢96 a]。本稿では、その処理方式を一部変更した箇所（図1）と、抽出された重要区間の単語認識（20万語相当の大語彙単語認識）の手法について報告する。

2. 音声要約・話題要約

通常、対話の音声・話題要約を行うには、以下の2つの方法が考えられる。

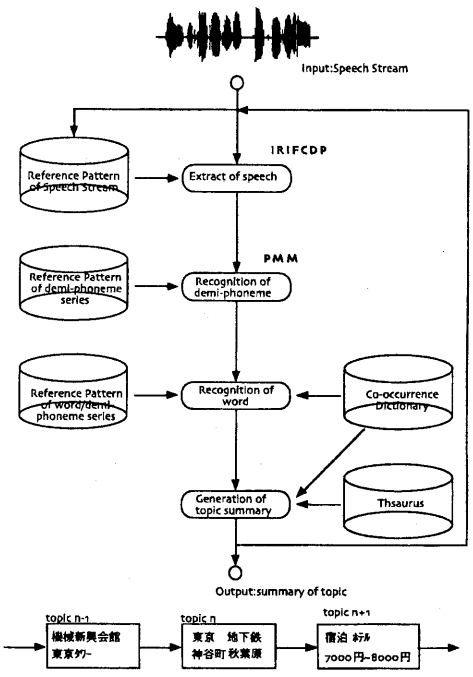

Figure 1 : Sequence of speech/topic summary system

(1) 自然対話の音声を全てを認識する
(2) 自然対話中に現れるその話題を反映した重要語句を抽出し、認識する

(1)の方法は、文認識・連続音声認識と呼ばれるものであり、一般的でまず第一に考えられる方法である。そして、厳密な意味・談話解析にとって必要かつ重要な方法である。しかし、この処理の過程で用いている文法は通常、書きことばを前提としており、話しことばに適応することは困難であると思われる。

話しことばの解析の難しさは、主語や述語、助詞の欠落、文の倒置等が原因となっている。この問題を解決するため、制約の緩い文法を構築して対処したとしても、これは誤った統語仮説候補を導き、結局は意味解析に負担をかけることになってしまう。実際そのような文法を構築することは可能であるのか、テキスト言語の場合と同様な能力を発揮できるのかという問題もある。それに加え、この方法はまず第一に文認識率が高いこと要求される。

本研究では、文法を用いた解析を行わず、音声中の互いに類似した十分な長さの区間という音声表層

の情報から、要約の生成を通して意味の深層情報を得るという従来とは異なったアプローチを試みている。(2) の方法である。このアプローチは、音声全体や文全体を理解するという深い意味や内容の理解には立ち入らずに、音声中に現れる単語の表層情報を手がかりにシソーラスなどの言語情報を用いて話題の推定を行うという、いわば斜め聞きといえるものである。

3. 話題の重要キーワード抽出

3.1 繰り返し出現する重要語句

ある話題において重要なキーワード、例えば固有名詞はその話題音声にしばしば出現し、普通名詞、助詞などのような単語に比べ継続時間が長いことが予想される。これは、その話題が固有に持っている事柄を発話を通して、繰り返し表現しているためだと思われる。

このような性質を利用し、音声要約に発展させた例が木山らによって提案されている[木山 95]。この方法は、音韻的に似ている区間を抽出し、その得られた区間と時間位置情報を用いて、話題のセグメンテーションと音声の要約を行うものである。この方式は、話題に依存せず、音声波形以外にも適応可能であり、事前に学習などの必要がないため、動作環境に依存しないというすぐれた特徴をもっている。しかし、スペクトル次元での特徴量を採用していることから、音声波形は同一話者に限られており、複数の話者の発話が含まれている場合には、本手法の適用は困難であると予想している。

要約に関する従来の研究の多くは、テキスト文字列を対象としたものである。文書の分類という点では、その文書を特徴づける単語を特定することなく、文書中に存在するそれぞれの単語の出現分野の推定をシソーラス、共起関係を使い分類を行った研究がある。[湯浅95ab] [小作 96, 亀田 96, 吉田 96, 大井 96]

このように、要約を目的として音声波形自体を対象としたものは少ない。しかし、これら論文の多くが、重要単語を繰り返し出現するキーワードと仮定しているものが多い。本研究でも同様な仮定を行って、重要単語の抽出を行っている。

本研究では、長時間の音声中の類似区間を抽出するのに適した手法として木山らによって提案されている Incremental Reference Interval-free 連続 DP (IRIFCDP) [木山95]を用いて、音声中の類似区間抽出を行ったのち、リアルタイム処理に適した手法として、岡によって提案されている連続 DP を用いた部分整合法 (Partial Matching Method ; PMM) [岡 86] を利用して音素片記号系列に変換する。

筆者による前回の報告 [中沢 96a] では、音素片記号系列から類似区間の抽出を行ったが、これを音声特徴量（スペクトル次元）に変更する。音素片記号系列の場合は、不特定話者の任意話題に対応した特徴量を重要単語の抽出に利用できるという利点があるが、実験の結果、スペクトル次元の特徴量に比べ、抽出率（再現率、適合率）の点において、著しく劣っているためである。(図2)

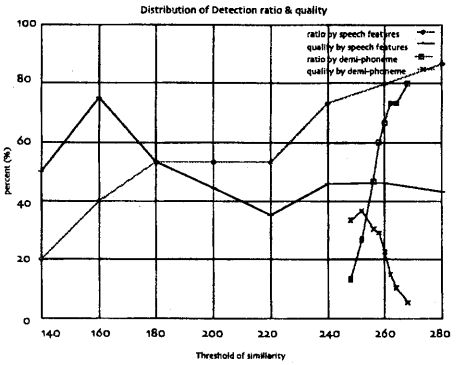

Figure 2: Distribution of Detection ratio & quality
(speech features & demi-phoneme)

3.2 重要キーワードの抽出

話題要約を表す自立語集合を抽出する第一近似として、音声表層情報を用いた重要キーワードの抽出を行う。キーワードモデルは次のようになる。

話題中に複数回出現する単語

重要区間集合 W は、抽出する類似区間の最小継続時間 *tmin*、探索領域 *tmax* によって得られる類似区間集合の結果に対し、パワーによる閾値 *power*、と継続区間長 *length* によって制約された集合とする。

$$W = \left\{ w_i \;\middle|\; (0 \le i \le N) \right\} \left| \begin{array}{c} w_i \in \mathrm{IRIFCDP}(tmin, tmax), \\ \mathrm{P}(w_i) \ge power, \\ \mathrm{L}(w_i) \ge length \end{array} \right.$$

Where, $\mathrm{P}(w_i)$ equal a power of w_i, $\mathrm{L}(w_i)$ equal a length of w_i

4. 音素片記号系列と大語彙単語認識

音声・話題要約システムは、入力音声として任意話題の対話を想定しているため、必然的に単語認識

は大語彙を対象としたものになる。これまでは、必要な単語を含む少語彙の標準パターンを用いて、連続DPにより単語認識（33単語、認識率81.56%）を行っていた。大語彙に対応するために、標準パターンの数をこのまま増やしたとしても、標準パターンの数（単語の数）に応じて線形に計算時間が増大するため、現実的ではない。

4.1 音素片を用いた大語彙単語認識の課題

音素片記号系列（複数の候補）を用いて単語認識を行う場合、通常は音素片記号の組み合わせによる単語辞書検索が行われる。しかし、この組み合わせは膨大なものとなり、少語彙ならともかく、大語彙の単語認識においては、計算時間の点で現実的には不可能になる。フレーム単位で音素片の認識結果を出力するようなシステムの場合、候補数が1つならば大きな問題はない。しかし、音素片候補がN個の場合、Mフレームの候補系列を検索するとなると、NのM乗回の検索が必要になる。認識結果を1つにした場合は認識率が落ち、認識率を上げるために複数候補にした場合には、今度は検索時間の問題にぶつかるというジレンマに陥る。

4.2 認識された音素片候補系列の特徴

このような問題に対処するため、まず音素片記号系列自体が持つ特徴を考える。実音声からフレームごとに切り出された区間の音素片記号系列には、次のような特徴があると考えられる。

(1) 継続区間内では同じ音素片が続いている
(2) 音素片の遷移が起こる場合は、前のフレームで認識されていた音素片との交差が起こる
(3) 各フレームにはいくつかの音素片候補がある
(4) 類似区間として得られた区間音素片の始端、終端は必ずしも単語の始端終端ではない

図3の例では、/ASA/という音素系列に関して、

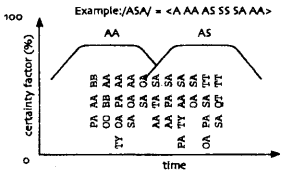

Figure 3: Transition of certainty factor

認識された音素片認識候補の遷移過程を示している。(1), (2)の現象にあるように最初の段階では、AAという音素片の確信度が高く数フレームに渡って出現するが、次第に確信度が低くなり、変わってASという音素片が現れるようになる。誤認識より確信度が前後した場合でも、(3)の情報から、この複数の認識候補から漏れることは少なくなり、音素片は連続して出現することになる。

```
D={d_i(t): 0≤t<T, 0≤i<N_t}
E={e_i^label, e_i^start, e_i^end : 0≤i<M}

SEARCH():
for(i=0;i<M;i++){
  add demi-phoneme label
  SEARCH_RECURSIVE( e_i^label, i+1, e_i^start, e_i^end )
  remove demi-phoneme label
}

SEARCH_RECURSIVE( e^label, j, e^start, e^end ):
for(i=j;i<M;i++){
  if(CROSS(e^start, e_i^start)){
    add demi-phoneme label
    if(CONNECT(demi_phoneme_labels,&index)){
      if(LENGTH(demi_phoneme_labels)-index>Preload){
        remove demi-phoneme label
      }
    }
    SEARCH_RECURSIVE(e_i^label, i+1, e_i^start, e_i^end)
    if(CONNECT(demi_phoneme_labels)){
      retreival in dictionary using demi_phoneme_labels
    }
    remove demi-phoneme label
  }
}

CROSS(i,j)            :if i cross j then 1 else 0
CONNECT(label,index)  :if label is connectable then 1 else 0
                        (index is set to rejected point)
LENTH(label)          :length of label
```
Algorithm1: Search for possible demi-phoneme series

4.3 交差・接続可能性のある音素片候補系列の利用

認識しようとする単語の音素片系列は、無秩序に組み合わされているわけではない。短い時間間隔では、声道・ピッチの変化がスペクトルの時間的変化を滑らかにしているため、ある規則に沿った音素片系列が認識される。一方、認識という立場から単語の表現方法を考えた場合、単語辞書には、見出し・読みがあり、読みを音素へ、音素を音素片へと変換したものが格納される。このように、各単語には、1対1、もしくは1対数組の対応関係が存在する。

そこで、得られた複数の音素片候補の系列の認識に、音素から音素片への変換規則を逆に利用し、音素として存在し得る音素片系列の推定を行い、組み合わせの数の削減を計ることを試みる。すなわち、単語認識する際に、音素片として接続可能な組み合わせだけを単語と仮定し検索しようというわけである。音素から音素片への変換の際に、ある特定の音

素片の脱落・挿入がある場合には、変換規則はネットワーク表現され、複数個の組み合わせを許すものなるだろう。しかし、このような場合でも、上記検索時の組み合わせの数への問題には十分に対処できると考えられる。(4) については、組み合わせる音素片の始端を自由（1フレームずつシフト）にすることで十分に対処できる。

音声区間 T（類似区間として抽出された区間）を音素片認識した結果である音素片記号系列のフレーム t における i 番目の音素片を $di(t)$ とする。今、この音素片 $di(t)$ が数フレームに渡り連続して出現したときの音素片ラベル、その開始・終了フレームをそれぞれ $e_i^{label}, e_i^{start}, e_i^{end}$ とする。そして、音素片の交差、接続可能条件を用いた単語認識は、アルゴリズム1のSEARCHで与えられる。*Preload* は複数の音素片が1つの音素に変換されることに対処するため、次のフレームの音素片を先読みするための値である。

5. 評価実験

この評価実験は、スペクトル次元の特徴量を採用して得られた類似区間とその区間の音素片記号系列の単語認識を目的としている。評価には、研究用データベースの話者 can0002 の発声した模擬対話セットKの30文を接続した音声（83.3秒）を対象とした。PMM を用いた音素片のフレーム認識率は、第1候補のみの場合、71.6% であり、第6候補まで含めた場合は、90.1% である。音声波形データは、フレーム間隔 8 msec で処理し、FFT 分析後 20 次元に圧縮されたスペクトルパラメータを特徴量として音素片に変換している。単語辞書には、EDR 辞書の日本語単語辞書を用いている。

- ●総レコード数　：386,803
- ●見出し異なり数：350,119
- ●読み異なり数　：233,611

すなわち、音素片記号系列を対象とした認識語彙数は、233,611 となる。尚、現段階では、音素から音素片への変換規則はネットワーク表現されてはいない。そのため、単語認識に用いる音素片記号系列は、脱落・挿入に関しては 100% の精度が求められる。

5.1 類似区間抽出

スペクトル次元の特徴量を用いた類似区間抽出と音素片を用いた類似区間抽出では、前者が抑揚などスペクトル次元で変化するものには、対応が難しいが、破裂音などの若干信号パワーが低い音声同士に対しては有効だという結果を得た。一方、音素片を特徴量とする類似区間抽出は、抑揚などの音声の変化にはかなり柔軟に対応するが、テンポがずれた音声、信号パワーが低い音声同士に対しては、音素片認識率事態が悪化するため、類似区間の抽出が難しいという結果を得た。（[中沢 96a] で報告済）

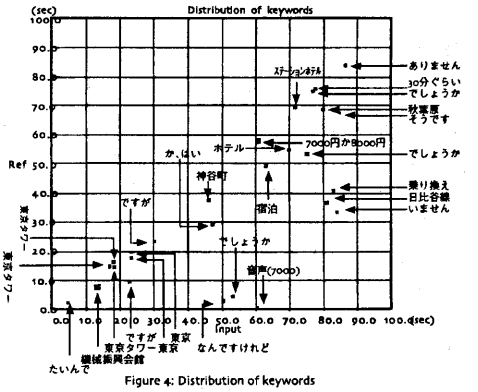

Figure 4: Distribution of keywords

5.2 単語認識

図5は、類似区間として認識されたもののうち「東京」という区間を音素片認識した実際の結果である。この音素片記号系列を用いて単語認識する際、その検索回数は、認識された53フレームの音素片認識結果を単純に連結して検索した場合、3.341874e39回(1)、連続している音素片を切出しその組み合わせで検索した場合、1.374390e11回(2)であると考えられ、交差及び接続可能なものの組み合わせで検索した場合は59回であった。検索回数は、(1), (2)それぞれに対し1.765477e-36%、4.292814e-8%にまで減少している。検索時間は、0.70秒であった。(1), (2)は時間がかかりすぎたため未計測である。尚、音声区間全て（83.3秒）中に含まれる単語を全て検索した場合は、3分44秒20であった。図6に連続して出現し交差している音素片の分布を、図8に検索結果を示した。この結果は、その後累積された確信度を元に最終的な単語の特定が行われる。

今回は、この手法を音素片に応用したが、読みから音素への変換規則も同様に、容易に適応できると考える。そして、この方法を利用すれば検索回数も現在より少なくなると予想される。

6 まとめ

音声・話題要約システムの音素片記号系列を対象とした大語彙の単語認識手法について、その有効性を述べた。この手法は、アルゴリズムが単純かつ高速であり、同時に並列化も可能であるという特徴をもつ。今後は、話題生成部分の評価実験を通して、本研究の有効性の証明と自然言語テキストで用いられる話題生成技術の利用、課題となっている同音意義語や未知語の処理、抽出区間の前後関係の探索を通して単語の同定方法の改善を計っていきたい。

謝辞

本研究の機会を与えて下さった新情報処理開発機構島田潤一所長に深く感謝致します。また、熱心に議論頂く情報統合研究室の皆様、音素片についてご指導下さる電総研 田中和世様に深く感謝致します。本研究では電子総合研究所の音声データベース "ETL - WD I-1" 及び日本電子化辞書研究所の EDR 辞書を使用致しました。

参考文献

[中沢 96a] 中沢正幸、古川清、豊浦潤、岡隆一：" 音声波形からの音素片記号系列を用いた音声要約と話題要約の検討"、電子情報通信学会技術研究報告書、SP96-28、p61-68

[木山 95] 木山次郎、伊藤慶明、岡隆一：" Incremental Reference Interval-free 連続DPを用いた任意話題音声の要約"、信学技報、SP95-35、(1995-06)

[岡 87] 岡隆一：" 連続DPを用いた部分整合法フレーム特徴の音韻認識"、電子情報通信学会誌、D Vol.J70-D No.5、pp.917-924 (1987-05)

[田中 86] 田中和世、速水悟、大田耕三：" 音声の音素片ネットワーク表現と時系列のセグメント化法を用いた自動ラベリング手法"、日本音響学会誌、42巻11号 pp. 860-868、(1986)

[小作 96] 小作浩美、伊佐原均：" 知的ニュースリーダにおける表層的話題関連性の抽出"、言語処理学会第2回年次大会論文集、pp93-96

[亀田 96] 亀田雅之：" 疑似キーワード相関法による重要キーワードと重要文の抽出"、言語処理学会第2回年次大会論文集、pp97-100

[吉田 96] 吉田和広、徳永健伸、田中穂積：" 新聞記事要約のためのテンプレートの自動抽出"、言語処理学会第2回年次大会論文集、pp105-108

[大井 96] 大井耕三、隅田英一郎、飯田仁、" 単語間の意味的類似度に基づく文書検索手法"、言語処理学会第2回年次大会論文集、pp109-112

[湯浅 95a] 湯浅夏樹、外川文雄、" 概念識別子の頻度分布を利用した文書分類"、情報学基礎 39-5、p33-40

[湯浅 95a] 湯浅夏樹、上田徹、外川文雄、" 大量文書データ中の単語共起を利用した文書分類"、情報処理学会論文誌、Vol 36 No8、p1819-1827

D	0	QK	QQ	QP	<QP	<QT	QT
D	1	QK	QQ	QP	<QP	<QT	QT
D	2	QK	QQ	QP	<QP	QT	<QT
D	3	QK	QQ	QP	<QP	QT	<QT
D	4	QK	QP	QQ	<QP	QT	<QT
D	5	QK	QP	QQ	<QP	QT	<QT
D	6	QK	QP	<QP	QQ	QT	<QT
D	7	QK	QP	<QP	QQ	QT	<QT
D	8	QK	QP	<QP	QQ	QT	<QT
D	9	KK	QP	PP	QK	TT	<BO
D	10	PP	KK	TT	QP	QK	<BO
D	11	KK	PP	TT	<BO	FO	
D	12	KK	PP	TO	<BO	FO	
D	13	KK	TO	PO	PP	<BO	FO
D	14	KK	TO	PO	PP	<BO	FO
D	15	KK	TO	PO	PP	<BO	FO
D	16	KK	TO	PO	MO	OO	FO
D	17	OOO	KK	OO	N'O	MO	PO
D	18	OO	OOO	KK	N'O	PO	TO
D	19	OO	OOO	N'O	KK	TO	PO
D	20	OO	OOO	N'O	KK	TO	ON'
D	21	OO	OOO	ON'	KK	TO	
D	22	OO	OOO	ON'	KK	OR	TO
D	23	OO	OOO	ON'	KK	OR	
D	24	OOO	OO	ON'	OG	OR	KK
D	25	OK	OP	ON'	OG	OB	OR
D	26	OK	OP	OG	ON'	OB	OR
D	27	OK	OP	OG	OB	OR	ON'
D	28	OK	OP	OG	OB	OR	
D	29	OK	OP	OG	OB	OR	
D	30	OK	OP	OG	BB		
D	31	OK	OP	QK	OG	BB	
D	32	QK	QP	OK	CCH	BB	OG
D	33	QK	QP	CCH	BB	OK	OG
D	34	PP	QK	KK	CCH	BB	OK
D	35	PP	QK	KK	CCH	BB	
D	36	PP	QK	KK	CCH		
D	37	PP	KK	QK	CCH		
D	38	KK	PP	CCH	QK	SHI	<Y
D	39	KK	PP	SHI	CHU	QK	<Y
D	40	KY	KK	IM	CHU	PP	QK
D	41	KY	KK	UU	IM	UR	PP
D	42	KY	UU	KK	MM	OR	UR
D	43	KY	NO	UU	MM	OR	UR
D	44	YO	RO	NO	MM	UU	OO
D	45	YO	RO	NO	OO	OOO	MM
D	46	YO	RO	NO	OO	OOO	MM
D	47	OO	YO	RO	OOO	ON	MM
D	48	OO	YO	OOO	RO	MU	ON
D	49	OO	OOO	MU	ON		
D	50	OO	OOO	MU			
D	51	OO	OT	OSH	UU	OG	
D	52	OO	OP	OG	OT	OSH	UU

Figure 5: Demi-phoneme series data

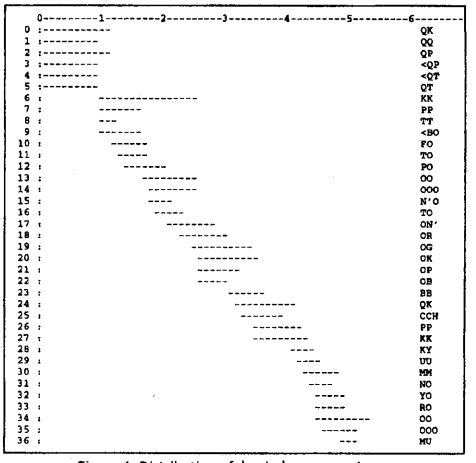

Figure 6: Distribution of demi-phoneme series

"水"	"PO"	"QP PP PO OO"
"ト"	"TO"	"QT TT TO OO"
"トウキョ"	"TO-KYO"	"QT TT TO OOO OK QK KK KY YO OO"
"トウキョウ"	"TO-KYO-"	"QT TT TO OOO OK QK KK KY YO OOO"
"トウ"	"TO-"	"QT TT TO OOO"
"キョ"	"KYO"	"QK KK KY YO OO"
"キョウ"	"KYO-"	"QK KK KY YO OOO"

Figure 8: Recognized word