概 要

音声合成の方法の 1 つとして提案されている音節波形接続方式 [1] は，既に録音された音声から，条件が一致する音節素片を切り出し，接続して音声を合成する．この方式は，信号処理を加えないで接続することにより，自然性の高い合成音声を作成できるという特徴がある．この方式の過去の研究としては，固有名詞 [2]，普通名詞 [3]，文節 [5] を対象として行われた．その結果，品質の高い合成音声が得られたことが報告されている．

しかし，問題点の 1 つとして，合成音声を作成する際に必要となる音節境界情報のラベルは，人手によって作成されるため，コストがかかる点が問題である．その解決方法の 1 つとして，音節境界情報の自動ラベリングが提案されている．そこで本研究では，音節波形接続方式において，自動ラベリングを使用して作成した合成音声の品質を調査した．

その結果，聴覚実験におけるオピニオン評価において，自動ラベルは 3.4，また，自動ラベルおよび手動ラベルを用いた合成音声の対比較実験において，自動ラベルは 40% という結果を得た．これより自動ラベルを用いた合成音声は，手動ラベルと差が小さく，品質の高い合成音声を作成できることが分かった．

目 次

図 目 次

表 目 次

1　はじめに

現在，カーナビゲーションシステムや電車の車内アナウンスなどのように，音声ガイダンスは様々な場面において利用されている．この音声ガイダンスの作成には，録音編集方式が広く使われている．録音編集方式というのは，ユーザーに依存しない比較的長い文音声 (以下，固定部) と，ユーザーに依存する比較的短い単語・文節音声 (以下，可変部) を別々に録音しておき，必要に応じて組み合わせることで，目的となる出力音声を作成する方式である．

例えば,「次の交差点を左折です.」という音声ガイダンスを作成したい場合,「次の○○左折です.」という固定部に,「交差点を」という可変部を挿入して作成する．

録音編集方式を用いた音声合成においては，固定部と可変部を接続した場合に違和感を軽減するために，一般に同一話者の音声を必要とする．固定部と可変部を別々に録音することにより，必要となる全ての音声を録音する場合に比べて，話者に対する負担は若干軽減されるが，可変部に挿入する音声が増大すると，大量の音声を同一話者から録音するのは困難となる．そこで，固定部は録音音声，可変部は合成音声を用いる方式がとられている．その合成音声を作成する方法の 1 つとして，音節波形接続方式 [1] が提案されている．

音節波形接続方式は，言語的なパラメータのみで合成音声を作成する方式であり，信号処理を加えないで接続することにより，自然性の高い合成音声を作成できるという特徴がある．この方式の過去の研究としては，固有名詞 [2]，普通名詞 [3]，文節 [5] を対象として行われた．その結果，品質の高い合成音声が得られたことが報告されている．

しかし，問題点の 1 つとして，合成音声を作成する際に必要となる音節境界情報のラベルは，人手によって作成されるため，コストがかかる点が問題である．その解決方法の 1 つとして，音節境界情報の自動ラベリングが提案されている．

そこで本研究では，音節波形接続方式において，自動ラベリングを使用して作成した合成音声の品質を調査する．なお，自動ラベリングの研究は，従来から多くの研究機関で行われており，HMM 法とベイズ確率を用いた統計的・確率的モデルによる方法 [8]，ルールベースを用いる方法 [9]，知識処理に基づく方法 [10] などが提案されている．

以降，2 章で音節波形接続方式を用いた音声合成について説明する．また，3 章で自動ラベリングについて説明し，4 章で実験環境と評価方法について説明する．そして，5 章で実験結果を示し，実験結果に対する考察を 6 章で述べる．

2　音節波形接続方式

2.1　音声合成に使用する素片

音節波形接続型音声合成は，波形編集型の音声合成方式の 1 種で，音響的なパラメータを使用しないで，言語的なパラメータのみで合成音声を作成することを特徴としている．具体的には，音節波形接続型音声合成では，既に録音された音声から，表 1 に示す条件が一致する音節素片を切り出し，接続して音声を合成する．

表 1: 実験に用いた音節素片

中心の音節
直前の音素 (前音素環境)
直後の音素 (後音素環境)
文節中のモーラ位置
文節のモーラ数
文節のアクセント型

音節波形接続型音声合成の例として「なっている (na-q/te/i/ru)」,「学校に (ga-q/ko-u/ni)」,「結婚した (ke-q/ko-N/shi/ta)」を音声合成する場合を以下に示す．なお,「_|¯」は音の強弱（アクセント）を表している．() 内強調部は，実際に選択される部分を示している．

なっている (na-q/te-i/ru)
　　＝なってから (**na-q**/te/ka/ra)
　　＋もっている (mo-q/**te-i**/ru)
　　＋増えている (fu-e/te-i/**ru**)
学校に (ga-q/ko-u/ni)
　　＝学校を (**ga-q**/ko-u/o)
　　＋実行に (ji-q/**ko-u**/ni)
　　＋ほんとうに (ho-N/to-u/**ni**)
結婚した (ke-q/ko-N/shi/ta)
　　＝結婚する (**ke-q**/ko-N/su/ru)
　　＋結婚して (ke-q/**ko-N**/shi/te)
　　＋逆転した (gya/ku/te-N/**shi**/ta)
　　＋結束した (ke-q/so/ku/shi/**ta**)

2.2 韻律，継続時間，調音結合の情報

一般に音声合成を行う場合，韻律の扱いが重要である．CHATR[7] などの通常の波形接続型音声合成では，ToBI モデルや藤崎モデルで韻律モデルを推定し，推定したピッチ周波数に類似した音素素片を，録音した音声データのなかから選択する．しかし，特定話者の単語発話を合成音声に使用する場合，単語のモーラ情報 (モーラ数とモーラ位置) が決まれば，単語によらずピッチ周波数がほぼ決定されることが知られている [2]．また，一般名詞の場合，名詞のモーラ情報に，名詞のアクセント型を加えることによって，非常に高い品質の音声が得られる [4]．文節発声の場合では，発話速度が遅い音声の場合には，文節単位でゆっくりと区切るためピッチが初期化される．それにより，文節発声の音声も一般名詞のみの発話と同様に扱うことができる [5].

音節波形接続型音声合成は，これらの事柄を利用して，韻律情報を，主に，モーラ数，モーラ位置，アクセント位置から得ている．また，音節継続時間の情報は，主に，音節の前後環境と，モーラ長およびモーラ位置から得ている．そして，調音結合の情報は，主に，音節の前後の環境から得ている．

2.3 波形接続に関する補則

音節波形接続型音声合成における波形接続に関する補則を以下に述べる．

2.3.1 連続母音

音節波形接続方式で作成された合成音声は，音声素片の接続部に違和感を生じる．特に違和感を生じる部分は，母音や撥音，促音が連続する部分である．これらの音節は前後の音が連続的に変化する部分であり，音節境界がはっきりせず，切り分けるのは困難である．これは，母音や撥音，促音が連続する場合，連続母音として扱うことにより，違和感を軽減できる [5].

2.3.2 音量，発話速度

音節波形接続方式では信号処理を加えないため，合成音声に使用する素片の音量の差が，音声の品質に直接影響する．そこで，音声の録音した時間帯が分かっている場合では，録音した時間帯が近い素片を選択し，音量や発話速度の均一化を行う [3].

2.3.3 音節素片の接続部

波形接続型音声合成では，接続部の違和感の発生が音声の自然性に大きく影響する．そのため，接続部における 2 素片間の波形の位相を考慮し，接続部の振幅の差がゼロに近づくように接続する．具体的には，あらかじめラベル付けされた素片開始時間と素片終了時間をもとに，振幅が負から正に変わる部分を，波形が短くなる方向 (開始時間は進む方向，終了時間は戻る方向) に探し，抽出する位置を修正する [1]．現在は，まだこの部分は人手による作業に依存している．

3　自動ラベリング

本研究では，自動ラベリングに HTK[13] を使用する．

3.1　HMM

HMM(Hidden Markov Model: 隠れマルコフモデル) は，状態遷移および出力シンボルが確率的に選択され，出力シンボルが与えられても状態遷移系列が唯一つに定まらない有限状態オートマトンである．HMM には，ある状態から全ての状態に遷移できる全遷移型 (Ergodic) モデルや，状態遷移が一定方向である left to right モデルなどがある．本研究では，left to right モデルを用いる．left to right モデルの例を図 1 に示す．

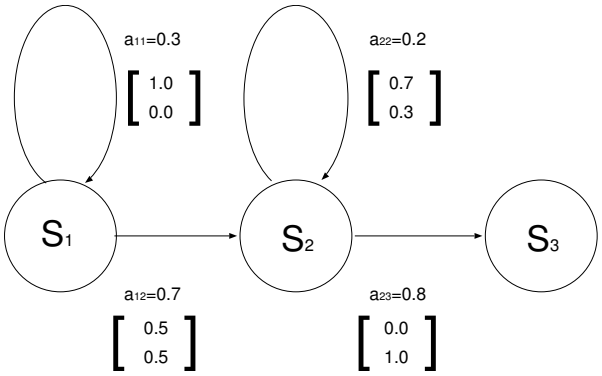

図 1: HMM(left to right モデル)

図 1 の HMM は，2 ループ 3 状態で構成され，2 種類のシンボル a，b を出力する．初期状態は S_1，最終状態 S_3 である．a_{ij} は状態 S_i から S_j に遷移する確率を示し，[] 内の数字はそれぞれシンボル a，b を出力する確率を示す．状態 S_1 を例にとると，状態 S_1 自身に 0.3 の確率で遷移し，1.0 の確率でシンボル a を出力し，0.0 の確率でシンボル b を出力する．

図 1 において，シンボル列 abb が与えられた場合，状態遷移系列は「$S_1 \rightarrow S_1 \rightarrow S_2 \rightarrow S_3$」,「$S_1 \rightarrow S_2 \rightarrow S_2 \rightarrow S_3$」の 2 通りが考えられ，状態遷移系列を唯一つに決定できない．よって，このモデルは隠れマルコフモデルといえる．

3.2 Baum-Welch アルゴリズム

Baum-Welch アルゴリズムは，与えられたシンボル列からモデルのパラメータを再推定するためのアルゴリズムである．HMM では，シンボル列を生成した状態遷移系列が観測できないため，直接，最尤推定を行うことができない．そのため，シンボル列に対する尤度を最大にするようにパラメータを再推定することを考える．Baum-Welch アルゴリズムについて，参考文献 [12] より引用し，以下に説明する．

与えられたシンボル列 $o_1^T = o_1 \cdots o_T$ に対し，時刻 t で状態 q_i から q_j に遷移した確率 $\gamma_t(i,j)$ は以下のようになる．

$$\begin{aligned}
\gamma_t(i,j) &= P(X_t = q_i, X_{t+1} = q_j \mid o_1^T, M) \\
&= \frac{P(X_t = q_i, X_{t+1} = q_j, o_1^T \mid M)}{P(o_1^T \mid M)} \\
&= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \alpha_T(i)}
\end{aligned}$$

また，時刻 t で状態 q_i に滞在した確率 $\gamma_t(i)$ を以下のように定義する．

$$\gamma_t(i) = \sum_{j=1}^{N} \gamma_t(i,j)$$

$\gamma_t(i,j)$ および $\gamma_t(i)$ を用いて，以下のようにパラメータを再推定できる．

$$\begin{aligned}
\bar{\pi} &= \gamma_1(i) \\
\bar{a}_{ij} &= \frac{\text{状態 } i \text{ から状態 } j \text{ に遷移する回数の期待値}}{\text{状態 } i \text{ から遷移する回数の期待値}} \\
&= \frac{\sum_{t=1}^{T-1} \gamma_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \\
\bar{b}_i(k) &= \frac{\text{状態 } i \text{ に滞在してシンボル } k \text{ を出力する回数の期待値}}{\text{状態 } i \text{ に滞在する回数の期待値}} \\
&= \frac{\sum_{t:o_t=k} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}
\end{aligned}$$

ここで，$\bar{\pi}, \bar{a}_{ij}, \bar{b}_i(k)$ は再推定したパラメータで，それぞれ，初期状態確率，状態遷移確率，記号出力確率である．

上記の再推定式によってパラメータを求める方法が，Baum-Welch アルゴリズムである．

3.3 Viterbi アルゴリズム

Viterbi アルゴリズムは，シンボル列 O を生成したモデル M における最適な状態遷移系列 (最適経路) と，その経路上での確率を求めるためのアルゴリズムであり，本実験では，HMM の初期モデルの作成，音節境界位置の計算に使われている．シンボル列 $o_1^T = o_1 ... o_T$ に対する最適経路は，状態遷移確率 $P(o_1^T, q_1^T|M)$ を最大にするような経路 $q_1^T = q_1 ... q_T$ である．Viterbi アルゴリズムの詳細について参考文献 [12] より引用し，説明する．

時刻 t において，状態 $q_i(1 \leq i \leq N)$ がシンボル列 $o_1^t = o_1, o_2, \cdots, o_T$ 生成する確率の最大値 $\delta(i)$ を求める．

$$\delta_t(i) = \max_{q_1^t} P(q_1^{t-1}, X_t = q_i, o_1^t | \ M)$$

$\delta_t(i)$ は，以下のように再帰的に計算できる．

$$\delta_{t+1}(i) = \max_i [\delta_t a_{ij}] b_j(o_{t+1})$$

ここで，a_{ij} は状態 q_i から q_j に遷移する確率，$b_i(o_t)$ は状態 q_i でシンボル o_t を出力する確率である．
また，状態遷移系列を復元するために，$\delta_{t+1}(i)$ を与える直前の状態 i を $\psi(\cdot)$ に記憶しておく．

$$\psi_{t+1}(j) = \arg\max_i [\delta_t(i) a_{ij}]$$

再帰計算が終了したら，時刻 $t = T$ における最適経路を，$t = T-1, ..., 1$ において，

$$\hat{q}_t = \psi_{t+1}(\hat{q}_{t+1})$$

により復元する．ただし，$\hat{q}_T = \arg\max_i \delta_T(i)$ である．

以上の操作により，最適経路 $q_1^T = \hat{q}_1 \cdots \hat{q}_T$ が求められる．

3.4 HMM を用いた自動ラベリング

自動ラベリングの手順を以下に示す.

1. 音声とその音声のラベルが与えられた学習データに対し，Viterbi アルゴリズムを用いて，音節 HMM の初期モデルを作成する.
2. Baum-Welch アルゴリズムを用いて，作成された音節 HMM の初期モデルの再推定を行う.
3. 再推定された音節 HMM と Viterbi アルゴリズムを用いて，音声合成を行う音声に対し，音節境界位置を計算して，音声のラベルを作成する.

4　評価実験

4.1　実験環境

4.1.1　音声合成に用いる音声データベース

合成音声の対象として，複数の電子辞書から重文複文を抽出した日英対訳の例文集の文を使用する．この例文集は機械翻訳を目的にしたものである．この例文集に収録されている 1,000 文を使用し，女性話者 (プロのナレーター) に，文節発声で遅く発声した音声を音声データベースとして用いる．この収録された音声データベースに対して自動ラベリングを行う．収録した音声発話の一部を表 2 に示す．なお，表中の"-"は文節の区切りであり，収録時にポーズを入れて収録した．

表 2: 収録した音声発話の一部

番号	文例
0001	これは-人々に-愛唱されている-古い-民謡の-一つです
0002	この背広に-合いそうな-ネクタイを-何本か-見せてください
0003	投手は-次の-打者に-セカンドフライを-打たせて-アウトにした

4.1.2　音声合成を行う音声

評価実験において，合成音声と自然音声を比較するために，音声データベース [4.1.1] 章に同一の発話内容が存在する音声を合成する．評価実験に使用する 100 文節の，モーラごとの文節数の割合を表 3 に示す．

表 3: 実験に使用する 100 文節中のモーラごとの内訳

モーラ数	文節数
4mora	17
5mora	70
6mora	13

これらの発話内容に対し，自動ラベルを用いた場合と手動ラベルを用いた場合の合成音声を作成する．作成した音声の例を表 4 に示す．

表 4: 作成した音声発話の一部

番号	作成音声
1	本性を
2	政界を
3	一生に

4.1.3 自動ラベリングのための学習データ

学習データとして，ATR の単語発話データベース Aset に収録されている，奇数番号データの女性話者 10 名 (1 話者につき 2,620 単語)，合計 26,200 単語を使用する．なお，自動ラベリングを行う音声合成に用いる音声データベース [4.1.1 章] の話者は，学習データの話者には含まれないため，不特定話者の自動ラベリングになる．

4.1.4 自動ラベリングに用いた音響パラメーター

自動ラベリングに用いた音響パラメータは，標準的な HTK の音響パラメータを使用する．そのパラメータを表 5 に示す．

表 5: 自動ラベリングに用いた音響パラメータ

基本周波数	16kHz
分析窓	Hamming 窓
分析窓長	25ms
フレーム周期	10ms
音響モデル	3 ループ 4 状態 (連続分布型)
stream 数	3
特徴	12 次 MFCC+Δ 12 次 MFCC
パラメータ	+対数パワー+Δ対数パワー (計 26 次)
	(母音・撥音・無音・連続母音)
連続型	MFCC 10, Δ MFCC 10,
HMM の	対数パワー 2, Δ対数パワー 2
初期モデル	(その他の音節・ue・uq)
混合分布数	MFCC 4, Δ MFCC 4,
	対数パワー 1, Δ対数パワー 1

4.2 評価方法

合成音声において，自動ラベルと手動ラベルの品質の差を調査するために，音節波形接続型で作成した合成音声の聴覚実験を行う．なお，評価した文節は各 100 文節である．

合成音声の評価には，音声研究に関わったことのない学生 5 名を対象に，オピニオン評価，対比較実験の 2 種類の実験を行う．

1. オピニオン評価

   音声の自然性を調べるために，オピニオン評価を行う．自然に聞こえた度合を 5 段階 (1 が最も不自然，5 が最も自然) で評価する．

2. 対比較実験

   作成した音声の評価のために，対比較実験を行う．対比較実験は，自然音声，手動ラベルを用いた合成音声，自動ラベルを用いた合成音声の 3 種類を使用し，以下の 3 回行う．

   A 自然音声-手動ラベル

   B 自然音声-自動ラベル

   C 手動ラベル-自動ラベル

   各組合せにおいて，同じ内容の文節発声の 2 種類の音声を連続して聴き，どちらの音声が自然に聞こえたかを判定する．

5　実験結果

5.1　自動ラベルと手動ラベルの差の例

「なっている (na-q/te-i/ru)」,「学校に (ga-q/ko-u/ni)」,「結婚した (ke-q/ko-N/shi/ta)」を作成するときに使用した音声と音節，および音節の自動ラベルと手動ラベルを表 6，表 7，表 8 に示す．また，それぞれの音声の，自動ラベルおよび手動ラベルを用いた合成音声，自然音声の波形を図 2〜図 10 に示す．

表 6: 「なっている」に使用する音節の自動ラベルと手動ラベル

音声	用いた音節	自動ラベル (ms)	手動ラベル (ms)
なってから (na-q/te-i/ru)	na-q	100〜480	100〜493
もっている (mo-q/te-i/ru)	te-i	490〜780	495〜780
増えている (fu-e/te-i/ru)	ru	760〜960	792〜952

この例では,「増えている」の「ru」の音節開始時間の差において，最大 32ms の差があるだけであり，ラベル全体の平均としては，自動ラベルと手動ラベルにあまり差がないことが分かる．

また，合成音声「なっている」のオピニオン評価は，手動ラベルでは 4.4，自動ラベルでも 4.4 と，同等の数値であった．

表 7: 「学校に」に使用する音節の自動ラベルと手動ラベル

音声	用いた音節	自動ラベル (ms)	手動ラベル (ms)
学校を (ga-q/ko-u/o)	ga-q	100〜500	100〜507
実行に (ji-q/ko-u/ni)	ko-u	470〜840	502〜868
ほんとうに (ho-N/to-u/ni)	ni	840〜1130	856〜1122

この例では,「実行に」の「ko-u」の音節開始時間の差において，最大 32ms の差があるだけであり，ラベル全体の平均としては，自動ラベルと手動ラベルにあまり差がないことが分かる．

また，合成音声「学校に」のオピニオン評価は，手動ラベルでは 3.8，自動ラベルでは 3.4 であった．

表 8: 「結婚した」に使用する音節の自動ラベルと手動ラベル

音声	用いた音節	自動ラベル (ms)	手動ラベル (ms)
結婚する (ke-q/ko-N/su/ru)	ke-q	90〜230	100〜431
結婚して (ke-q/ko-N/shi/te)	ko-N	380〜690	415〜687
逆転した (gya/ku/te-N/shi/ta)	shi	730〜920	731〜989
結束した (ke-q/so/ku/shi/ta)	ta	910〜1110	956〜1111

この例では,「結婚する」の「ke-q」の音節終了時間の差において，最大 201ms の差があり，自動ラベルと手動ラベルに大きな差があることが分かる.

また，合成音声「結婚した」のオピニオン評価は，手動ラベルでは 3.8，自動ラベルでは 1.4 であった.

「なっている (na-q/te-i/ru)」の音声波形を以下に示す．自動ラベルを用いた合成音声を図 2，手動ラベルを用いた合成音声を図 3，自然音声の波形を図 4 に示す．

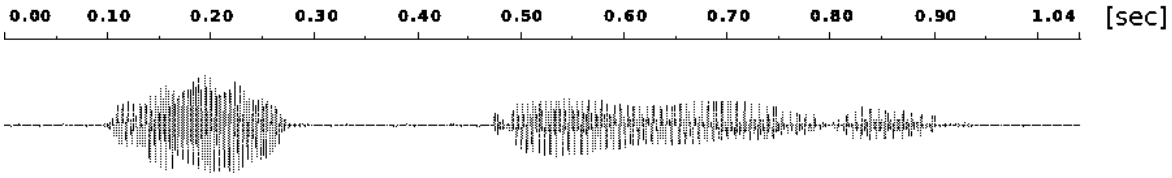

図 2: 自動ラベル「なっている」

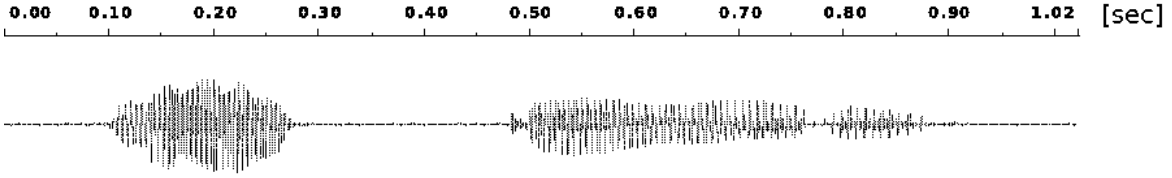

図 3: 手動ラベル「なっている」

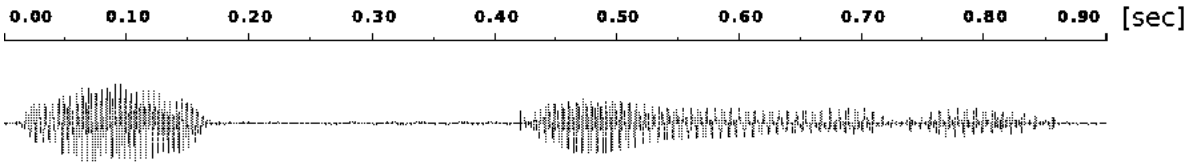

図 4: 自然音声「なっている」

図 2，図 3，図 4 より,「なっている na-q/te-i/ru」における，自動ラベルおよび手動ラベルを用いた合成音声，自然音声の音声波形にはそれほど大きな違いはないことが分かる．

「学校に (ga-q/ko-u/ni)」の音声波形を以下に示す．自動ラベルを用いた合成音声を図 5，手動ラベルを用いた合成音声を図 6，自然音声の波形を図 7 に示す．

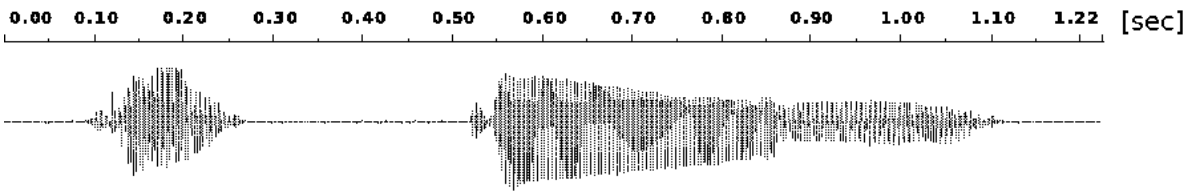

図 5: 自動ラベル「学校に」

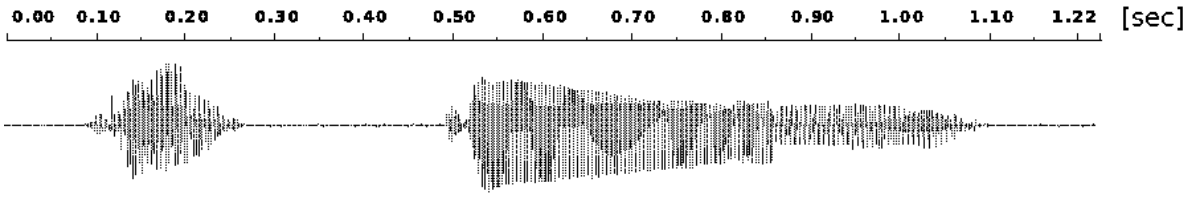

図 6: 手動ラベル「学校に」

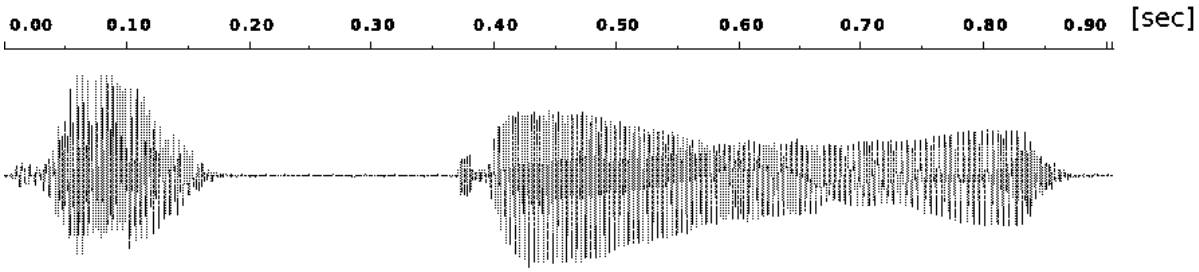

図 7: 自然音声「学校に」

図 5，図 6，図 7 より,「学校に ga-q/ko-u/ni」における，自動ラベルおよび手動ラベルを用いた合成音声の音声波形にはそれほど大きな違いはないことが分かる．しかし，合成音声と自然音声を比較すると，合成音声の「ko-u」と ni」の接続部において，若干ではあるが，音量の違いがあることが分かる．

「結婚した (ke-q/ko-N/shi/ta)」の音声波形を以下に示す．自動ラベルを用いた合成音声を図 8，手動ラベルを用いた合成音声を図 9，自然音声の波形を図 10 に示す．

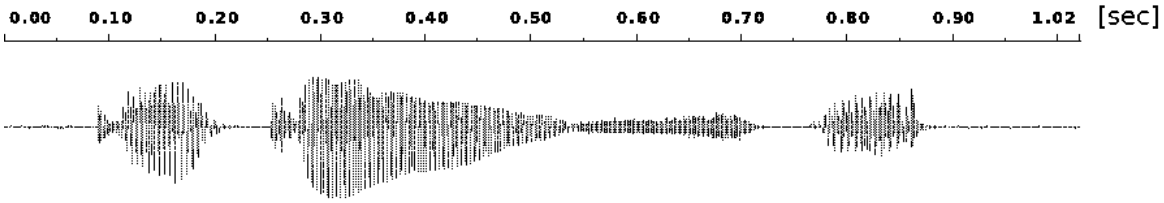

図 8: 自動ラベル「結婚した」

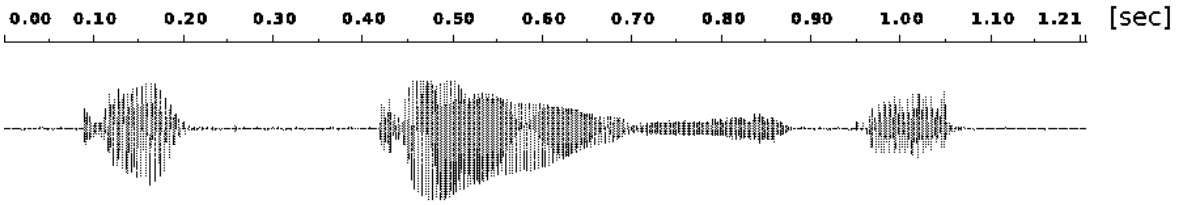

図 9: 手動ラベル「結婚した」

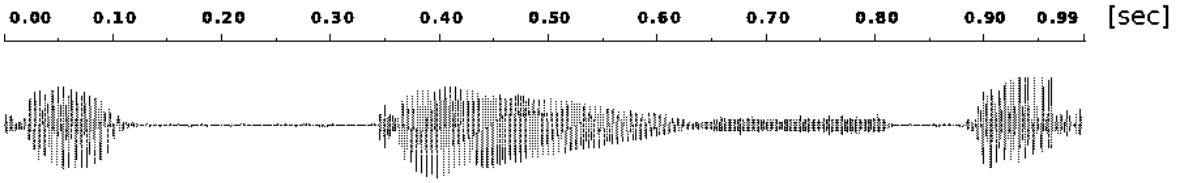

図 10: 自然音声「結婚した」

図 8，図 9，図 10 より,「結婚した ke-q/ko-N/shi/ta」において，自動ラベルおよび手動ラベルを用いた合成音声とを比較すると，自動ラベルの「ke-q」の音節継続時間が手動ラベルに比べ，短いことが分かる．

5.2 自動ラベルと手動ラベルの平均の差

自動ラベルと手動ラベルの音節開始時間，音節終了時間，および音節継続時間の差の分布図と結果を以降に示す．なお，評価した音節は各 338 音節である．

5.2.1 音節開始時間の実験結果

自動ラベルと手動ラベルの音節開始時間の差の分布図を図 11 に，平均値と標準偏差を表 9 に示す．

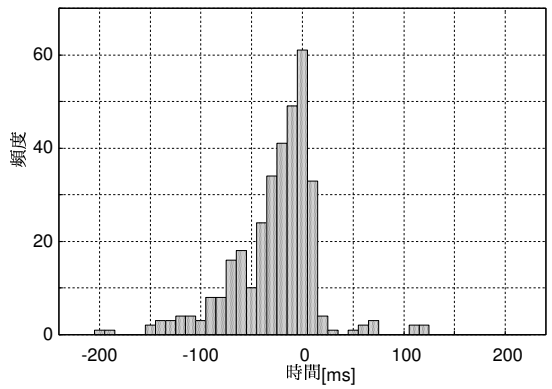

図 11: 音節開始時間の差の分布図 (自動ラベル-手動ラベル)

表 9: 音節開始時間における自動ラベルと手動ラベルの差

	平均値 (ms)	標準偏差 (ms)
音節開始時間	-26.6	42.0

表 9 より，音節開始時間において，自動ラベルは手動ラベルに比べ，約 25ms 早めにラベリングする傾向があることが分かった．

また，図 11 より，自動ラベルと手動ラベルの差の頻度が最大となる時間は 0ms 付近にあることが分かった．この時間に該当するデータは 61 音節ある．

図 11 において，自動ラベルと手動ラベルで大きく異なったデータを表 10 に示す．

表 10: 音節開始時間の大きく異なったデータ (自動ラベル-手動ラベル)

音声	用いた音節	差分 (ms)		音声	用いた音節	差分 (ms)
歩いていた	ta	-199		肝臓を	o	117
会場は	wa	-187		寸法を	o	116
返していた	ta	-146		同情を	o	112
強盗の	to-u	-146		健康を	o	107
文学を	ga	-143		友情を	o	74
燃えていた	ke	-138		学校を	o	73

音素別に調査した結果，音節開始時間の差において，自動ラベリングが早いデータとしては，中心の音節に無声破裂音「p,t,k」を含むものが多く存在した.

5.2.2 音節終了時間の実験結果

自動ラベルと手動ラベルの音節終了時間の差の分布図を図 12 に，平均値と標準偏差を表 11 に示す．

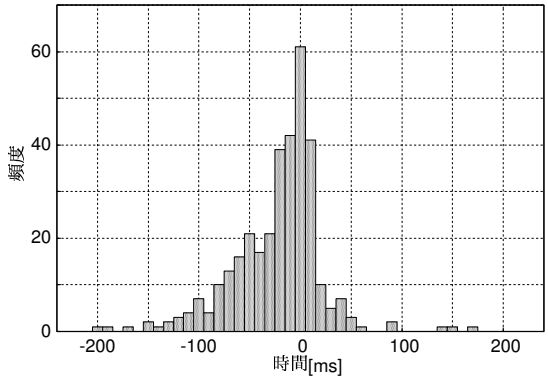

図 12: 音節終了時間の差の分布図 (自動ラベル-手動ラベル)

表 11: 音節終了時間における自動ラベルと手動ラベルの差

	平均値 (ms)	標準偏差 (ms)
音節開始時間	-22.6	43.8

表 11 より，音節終了時間においても，音節開始時間と同様に，自動ラベルは手動ラベルに比べ，約 25ms 早めにラベリングする傾向があることが分かった．

また，図 12 より，自動ラベルと手動ラベルの差の頻度が最大となる時間は 0ms 付近にあることが分かった．この時間に該当するデータは 61 音節ある．

図 12 において，自動ラベルと手動ラベルで大きく異なったデータを表 12 に示す．

表 12: 音節終了時間の大きく異なったデータ (自動ラベル-手動ラベル)

音声	用いた音節	差分 (ms)		音声	用いた音節	差分 (ms)
結婚する	ke-q	-201		復興を	ko-u	168
一生の	i-q	-186		文章を	sho-u	147
迫っていた	te-i	-166		本性を	sho-u	137
音楽は	o-N	-154		ほんとうに	ni	89
直っていた	te-i	-153		牧場に	ku	85
先方が	se-N	-137		政権を	ke-N	55

音素別に調査した結果，音節終了時間の差において，自動ラベリングが早いデータとしては，中心の音節に連続母音「e-i」を含むものが多く存在した．

5.2.3 音節継続時間の実験結果

自動ラベルと手動ラベルの音節継続時間の差の分布図を図 13 に，平均値と標準偏差を表 13 に示す.

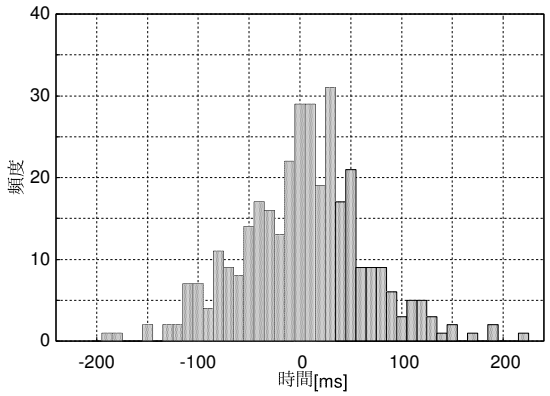

図 13: 音節継続時間の差の分布図 (自動ラベル-手動ラベル)

表 13: 音節継続時間における自動ラベルと手動ラベルの差

	平均値 (ms)	標準偏差 (ms)
音節開始時間	3.9	62.0

表 13 および音節開始時間，音節終了時間ともに約 25ms 早めにラベリングする傾向があることより，音節継続時間において，自動ラベルと手動ラベルの差は，ほとんどないことが分かった.

図 13 において，自動ラベルと手動ラベルで大きく異なったデータを表 14 に示す.

表 14: 音節継続時間の大きく異なったデータ (自動ラベル-手動ラベル)

音声	用いた音節	差分 (ms)		音声	用いた音節	差分 (ms)
結婚する	ke-q	-191		歩いていた	ta	217
一生の	i-q	-176		政権を	ke-N	188
音楽は	o-N	-154		復興を	ko-u	185
先方が	se-N	-147		会場は	wa	168
迫っていた	te-i	-129		文章を	sho-u	150
寸法を	o	-129		返していた	ta	149

5.3 オピニオン評価の実験結果

オピニオン評価の全被験者の平均を表 15 に示す.

表 15: オピニオン評価の実験結果

	オピニオン評価
自動ラベル	3.4
手動ラベル	3.7
自然音声	4.6

表 15 より，自動ラベルを用いた合成音声は，手動ラベルを用いた合成音声より品質がやや低いが，差が小さいことが分かった．また，自動ラベルを用いた合成音声は，自然音声には及ばないものの，オピニオン評価において 3.4 という高い値を得た.

5.3.1 自動ラベルで良い評価となった音声

自動ラベルを用いた合成音声のオピニオン評価において，最も良い評価となった音声とオピニオン評価の数値を表 16 に示す．また，比較のために，手動ラベルを用いた合成音声のオピニオン評価の数値を示す.

表 16: 自動ラベルで良い評価となった音声

音声	自動ラベル	手動ラベル
部長の (bu/cho-u/no)	4.8	4.6
音楽に (o-N/ga/ku/ni)	4.8	4.4
仕事の (shi/go/to/no)	4.8	4.2
健康の (ke-N/ko-u/no)	4.6	4.2
遺体を (i/ta-i/o)	4.6	3.6

表 16 より，自動ラベルを用いた合成音声でも，品質の高い合成音声を得られることが分かった.

5.3.2 自動ラベルで悪い評価となった音声

自動ラベルを用いた合成音声のオピニオン評価において，最も悪い評価となった音声とオピニオン評価の数値を表 17 に示す．また，比較のために，手動ラベルを用いた合成音声のオピニオン評価の数値も示す．

表 17: 自動ラベルで悪い評価となった音声

音声	自動ラベル	手動ラベル
結婚した (ke-q/ko-N/shi/ta)	1.4	3.8
成功は (se-i/ko-u/wa)	1.4	2.4
優勝を (yu-u/sho-u/o)	1.8	3.6
大量の (ta-i/ryo-u/no)	1.8	2.4
政権を (se-i/ke-N/o)	2.0	1.8

表 17 より，自動ラベルを用いた合成音声は，手動ラベルに比べ，同程度またはそれ以下の品質になる音声があることが分かった．

5.3.3 自動ラベルが良い評価となった音声 (自動ラベル-手動ラベル)

自動ラベルおよび手動ラベルを用いた合成音声のオピニオン評価の差において，自動ラベルを用いた合成音声が最も良い評価となった音声を表 18 に示す．また，自動ラベルおよび手動ラベルを用いた合成音声のオピニオン評価の数値と，その評価の差を示す．

表 18: 自動ラベルが良い評価となった音声 (自動ラベル-手動ラベル)

音声	評価の差	自動ラベル	手動ラベル
そのような (so/no/yo-u/na)	1.6	3.8	2.2
遺体を (i/ta-i/o)	1.0	4.6	3.6
信条に (shi-N/jo-u/ni)	1.0	5.0	4.0
ありそうな (a/ri/so-u/na)	1.0	3.0	2.0
友情を (yu-u/jo-u/o)	0.8	4.0	3.2

表 18 より，自動ラベルを用いても，手動ラベルよりも品質の高い合成音声を作成できることが分かった．

5.3.4 自動ラベルが悪い評価となった音声 (自動ラベル-手動ラベル)

自動ラベルおよび手動ラベルを用いた合成音声のオピニオン評価の差において，自動ラベルを用いた合成音声が最も悪い評価となった音声を表 19 に示す．また，自動ラベルおよび手動ラベルを用いた合成音声のオピニオン評価の数値と，その評価の差を示す．

表 19: 自動ラベルが悪い評価となった音声 (自動ラベル-手動ラベル)

音声	評価の差	自動ラベル	手動ラベル
結婚した (ke-q/ko-N/shi/ta)	-2.4	1.4	3.8
大量に (ta-i/ryo-u/ni)	-2.4	2.4	4.8
膨大な (bo-u/da-i/na)	-1.8	2.8	4.6
優勝を (yu-u/sho-u/o)	-1.8	1.8	3.6
方法を (ho-u/ho-u/o)	-1.4	2.0	3.4

表 19 より，自動ラベルを用いた合成音声は，手動ラベルと比較すると，品質が劣化し，表 18 の評価の差よりも大きな差があることが分かった

5.3.5 評価に差がなかった音声 (自動ラベル-手動ラベル)

自動ラベルおよび手動ラベルを用いた合成音声のオピニオン評価の差において，差がなかった音声を表 20 に示す．また，自動ラベルおよび手動ラベルを用いた合成音声のオピニオン評価の数値と，その評価の差を示す．

表 20: 評価に差がなかった音声 (自動ラベル-手動ラベル)

音声	評価の差	自動ラベル	手動ラベル
私の (wa/ta/shi/no)	0.0	3.2	3.2
なっている (na-q/te-i/ru)	0.0	4.4	4.4
研究を (ke-N/kyu-u/o)	0.0	3.0	3.0
小説を (sho-u/se/tsu-o)	0.0	3.8	3.8
詳細を (sho-u/sa-i/o)	0.0	2.8	2.8

20 より，自動ラベルおよび手動ラベルを用いた合成音声には，評価に差がないものがあり，それぞれの音声のオピニオン評価の数値は一様でないことが分かった．

5.4 対比較実験の実験結果

5.4.1 自動ラベルおよび手動ラベルを用いた合成音声の対比較

自動ラベルおよび手動ラベルを用いた合成音声の対比較実験の結果を表 21 に示す.

表 21: 自動ラベルと手動ラベルの対比較実験の結果

	自動ラベル (%)	手動ラベル (%)
文節数 100	40	60

表 21 より，自動ラベルを用いた合成音声は，手動ラベルを用いた合成音声より品質がやや低いが，差が小さいことが分かった.

5.4.2 自然音声との対比較

自動ラベルおよび手動ラベルを用いた合成音声のそれぞれと自然音声との対比較実験の結果を表 22 に示す.

表 22: 自然音声との対比較実験の結果

	自然音声 (%)	自動ラベル (%)
文節数 100	85	15
	自然音声 (%)	手動ラベル (%)
文節数 100	80	20

表 22 より，自然音声と合成音声を比較すると，合成音声の値は，一見低い値に見えるが，自然音声よりも自然と判定された音声が，自動ラベルでは 15%，手動ラベルでは 20%もあり，かなりよい数値を得た.

6　考察

6.1　音節境界位置の解析

表 9 と表 11 より，音節開始時間および音節終了時間において，自動ラベルは手動ラベルに比べ，約 25ms 早めにラベリングする傾向があることが分かった．しかし，図 11，図 12 より，自動ラベルと手動ラベルの差の頻度が最大となる時間は 0ms 付近にあることから，全体の自動ラベルに対し，補正を加えるべきではないと考えている．

しかし，表 10，表 12，表 14 から得られた結果より，自動ラベルと手動ラベルの差が大きい特定の音節 (無声摩擦音「p,t,k」，連続母音「e-i」を含む音節) に対して，何 ms の補正を加えるといったルールを追加することで，手動ラベルに近付く可能性はあると考えている．

また，本研究における不特定話者と従来の研究 [11] における特定話者の，自動ラベルと手動ラベルの音節境界位置および音節継続時間の差の平均値と標準偏差の結果を表 23，表 24 に示す．

表 23: 不特定話者の結果

	平均値 (ms)	標準偏差 (ms)
音節境界位置	-25	43
音節継続時間	-3.9	62

表 24: 特定話者の結果

	平均値 (ms)	標準偏差 (ms)
音節境界位置	0.50	29.49
音節継続時間	-2.71	42.65

表 23 と表 24 より，標準偏差の値に注目すると，音節境界位置および音節継続時間ともに，不特定話者の標準偏差は特定話者の標準偏差の約 1.5 倍大きくなっていることが分かった．しかし，聴覚実験の結果から，不特定話者の自動ラベリングでも十分な品質が得られた．したがって，コストのかかる特定話者の自動ラベリングをする必要はないと考えている．

6.2　オピニオン評価および対比較実験の解析

オピニオン評価および対比較実験の結果から，手動ラベルと自動ラベルを用いた合成音声とを比較すると，自動ラベルを用いた合成音声の品質が大きく低下および向上した音声があった．

その原因としては，各音節のラベリングの精度の問題にあり，低下した理由は，音節の欠如，または音節の過多によるものと考えている．しかし，向上した理由は，手動ラベルも間違ってラベル付与されているわけではなく，明らかではない．

また，オピニオン評価および対比較実験の結果から，合成音声は，自然性の面で，自然音声に及ばないことが分かった．これは，合成に用いた録音音声の音量，音質の違いによるものと考えている．

6.3　本実験の信頼性

本実験の信頼性を確認するため，音節波形接続型音声合成を文節に適用した研究 [6] で示されたオピニオン評価の実験結果を表 25 に示す．

表 25: 先行研究のオピニオン評価の実験結果

	オピニオン評価
自然音声	4.8
手動ラベル	3.8

表 15 と比較すると，本研究のオピニオン評価は，先行研究より低いことが分かった．これは，非試験者がが同一でないため低くなったと考えている．

しかし，先行研究における，自然音声との差は 0.2，手動ラベルを用いた合成音声との差は 0.1 となり，ともに同程度，オピニオン評価が低くなっている．これより，本実験は先行研究と同等の信頼性が得られた実験であると考えている．

6.4　自動ラベルを用いた合成音声の考察

今回の実験では，手動ラベルと自動ラベルで，音節境界位置が大きく異なるデータが少しあった．しかし，オピニオン評価および対比較実験の実験結果から，自動ラベルと手動ラベルを用いた合成音声の品質の差は小さいことが分かった．これより，自動ラベルを用いた合成音声は，十分実用可能な音声であると考えている

7　まとめ

本研究では，文節発声の音声に対して，自動ラベリングを使用し，音節波形接続方式で作成した合成音声の品質をを調査した．聴覚実験の結果において，自動ラベルのオピニオン評価は 3.4，また，自動ラベルおよび手動ラベルの対比較実験において，自動ラベルは 40%という結果を得た．これより，自動ラベルを用いた合成音声は，手動ラベルとの差が小さく，品質の高い合成音声が作成できることが分かった．

今後の課題としては，より精度を高めるために，自動ラベルと手動ラベルの差が大きかった特定の音節に対してルールを追加する方法，また，人間の評価の低かった音声を調査し，対策することが考えられる．

謝辞

最後に，一年間に渡って御指導，御教授して頂きました鳥取大学工学部知能情報工学科計算機 C 研究室の池原教授と村上助教授に深くお礼申し上げます．加えて，本論文を執筆するにあたり，参考にさせて頂いた論文，聴覚実験，そして実験結果の集計，雑用，その他諸等に協力してくださった岡本一輝さん，片山慶一郎さん，松浦祥悟さん，植村和久さんに深く感謝いたします．

参考文献

[1] 村上仁一, 水澤紀子, 東田正信: “音節波形接続方式による単語音声合成”, 電子情報通信学会論文誌, J85-D-II, pp.1157-1165, (2002).

[2] 水澤 紀子, 村上 仁一, 東田 正信: “モーラ数，モーラ位置に基づいた音節波形接続による単語音声合成”, 日本音響学会論文集, 2-Q-16, pp.311-312, (1999).

[3] 石田隆浩, 村上仁一, 池原悟: “音節波形接続型音声合成の普通名詞への応用, 電子情報通信学会技術研究報告, SP2002-25, pp.7-12, (2002).

[4] 石田隆浩, 村上仁一, 池原悟: “モーラ情報とアクセント情報を用いた波形接続型音声合成の普通名詞への応用”, 日本音響学会 2003 年春季研究発表会, 2-Q-18, pp.409-410, (2003).

[5] 加藤琢也, 村上仁一, 池原悟: “波形接続型音声合成の文節への適用”, 日本音響学会 2004 年秋期研究発表会, 3-2-12, pp.339-340, (2004).

[6] 村上 仁一, 加藤 琢也, 池原 悟, “音節波形接続型音声合成の文節への適用”, 電子情報通信学会技術研究報告, SP2005-19, pp.43-50, (2005).

[7] Nich Campbell, Alan W.Black: “CHATR 自然音声波形接続型任意音声合成システム”, 電子情報通信学会技術研究報告, SP96-7, pp.45-52, (1996).

[8] 中川, 橋本: “HMM 法とベイズ確率を用いた連続音声のセグメンテーション” 電子情報通信学会論文誌, J72-D-II, pp.1-10, (1989).

[9] 古市, 相澤, 井上, 今井: “音声認識におけるルールベース法による話者独立音素セグメンテーション”, 日本音響学会誌, 55, pp.707-716, (1999).

[10] 鬼山, 荒井, 山下, 北橋, 野村, 溝口: “知識処理に基づく音声自動ラベリングシステム”, 電子情報通信学会技術研究報告, SP90-84, pp.53-60, (1991).

[11] 前田, 村上, 池原: “モーラ情報を用いた音素ラベリング方式の検討”, 電子情報通信学会技術研究報告, SP2001-53, pp.25-30, (2001).

[12] 北 研二: “言語と計算-4 確率的言語モデル”, 東京大学出版会, ISBN4-13-065404-7, (1999).

[13] Hidden Markov Model Toolkit(HTK) http://htk.eng. cam.ac.uk/