論文 / 著書情報
Article / Book Information

論題(和文)	立体的音像定位を利用した音声合成システムの検討
Title(English)	
著者(和文)	神山 歩相名, 岩野 公司, 飯田 一博, 古井貞熙
Authors(English)	Hosana Kamiyama, Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会2009年春季講演論文集, , , pp. 1513-1514
Citation(English)	, , , pp. 1513-1514
発行日 / Pub. date	2009, 3

立体的音像定位を利用した音声合成システムの検討*

◎神山歩相名（東工大），岩野公司（武蔵工大），飯田一博（千葉工大），古井貞熙 (東工大)

1　はじめに

最近は複雑なレイアウトの Web ページが増えており，文字の表示位置などのレイアウトに関する情報が重要な意味を持つようになりつつある．しかし，これまでのスクリーンリーダーや音声ブラウザは，画面上の文字や操作方法を読み上げる機能のみを実装したものがほとんどであり，視覚障害者にとっては，Web ページのレイアウト情報の把握が未だ難しい状況である．特に，糖尿病や交通事故などによる中途失明者は，晴眼時と同様にレイアウト情報を把握したいという願望が強いこともあり，このような情報の提示機能のついた視覚障害者向けの音声ブラウザの開発が望まれる．例えば，IBM 社の「ホームページ・リーダー [1]」では，Web ページの読み上げ箇所のテキスト部を反転させて，文字位置の提示を行っているが，弱視者を対象とした機能であり，全盲の視覚障害者にとっては有用性が乏しい．

そこで本研究では，Web ブラウザ上に表示された文章を，その文字位置に対応する方向から聞こえるように立体的に音像定位させて読み上げを行うことで，レイアウトに関する情報を合成音声に含めて提示するシステムを提案する．提案システムの合成音の音像定位感と読み上げ音声の移動の滑らかさについて主観評価を行った．またシステムの使用感の評価のため，音像定位のある場合とない場合の音声の聞き取りやすさと，システムの利便性について比較を行った．

2　システムの基本設計

本研究で提案するシステムの概要図を Fig.1 に示す．本システムは，「文字情報抽出部」「音声合成部」「音像定位制御部」から構成されている．以下，それぞれの機能を説明する．

2.1　文字情報抽出部

文字情報抽出部は，Web ブラウザ中で読み上げるべきテキストとその文字位置情報を抽出する部分である．文字位置情報は，読み上げテキストの各文字の Web ブラウザ上の座標として抽出される．文字情報抽出部は，Mozilla Firefox (ver 2.0) の拡張機能として実現した．

2.2　音声合成部

音声合成部は，文字情報抽出部から抽出されたテキストに対応する音声を，当研究室にて開発した TTS システムを用いて合成する [2][3]．TTS システムは，HMM 音声合成であり，ケプストラム情報をスキップのない 3 状態混合数 1 の left-to-right 型の triphone HMM としてモデル化する．

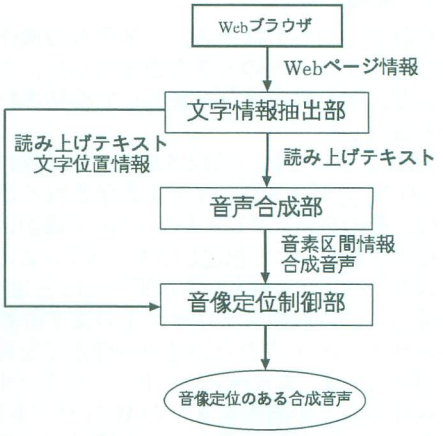

Fig. 1　Web ページ読み上げシステムの概念図

本研究では，HMM は ATR 日本語音声データベース中の男性話者 1 名 (MHT) による 450 発声 (A〜I セット各 50 発声) によって学習した．特徴量は，0〜25 次元のメルケプストラムとその Δ 係数である．メルケプストラムは，窓幅 16ms の Blackman 窓を用い，フレーム周期は 5ms で分析したスペクトルより求めた．音素長と F_0 は数量化 I 類によりモデル化を行った [2][3]．

2.3　音像定位制御部

音像定位制御部は，合成音の各音素区間に対して，文字位置に対応した頭部伝達関数 (HRTF) をたたみ込み，立体的音像定位のある合成音を生成する．HRTF は，半無響室で実際の頭部を用いて計測したものを用いた．スピーカーと被験者が正対する角度を基準 (0°) として，水平面上を 30° ごとに 12 点測定した．インパルス応答は，標本化周波数 48kHz で記録した．すべての文字位置に対応する方向の音像定位を実現するためには，測定した HRTF から，任意の角度の HRTF を推定する必要がある．この HRTF の推定には，西野らのインパルス応答を線形 2 点補間する手法 [4] を用いた．今回は，仰角方向の音像定位は行わず，水平方向の音像定位のみ行った．

3　評価実験

3.1　実験条件

システムを稼働させる PC は，CPU が Intel Core 2 Duo (3.00GHz)，内蔵メモリが 2.0GB であり，OS は Linux (Fedora 8) となっている．

HRTF は，大学院生 1 名のものを測定し使用した．受聴に用いるヘッドフォンには，audio-technica 社製

*A speech synthesizer using binaural sound localization by Hosana Kamiyama (Tokyo Institute of Technology), Koji Iwano (Musashi Institute of Technology), Kazuhiro Iida (Chiba Institute of Technology) and Sadaoki Furui (Tokyo Institute of Technology)

の ATH-M30 を用いた．実験は研究室内で行い，被験者は大学（院）生 18 名であり，全員晴眼者である．評価を行う被験者の中に，HRTF を測定した被験者は含まれない．

3.2 実験手順

被験者に対しては予め，システムの操作方法と実験で想定する画面の大きさを説明した．また本システムが，視覚障害者を対象として構築されていることも説明した．

Web ブラウザは，被験者から 63cm 離れた横縦比 4:3 の 20 型ディスプレイ上に表示されるが，被験者には，そのディスプレイが 4×3m の仮想スクリーンに拡大されたものと想定してもらった．このとき，仮想スクリーンは被験者から水平方向に $\pm 72°$ の範囲に位置し，この仮想スクリーン上の文字位置に対応するように HRTF をたたみ込み音像定位を実現した．

被験者には，ニュースサイト／ブログ／オンライン百科辞典などの適当な 5 つの Web サイトについて，音像定位のある合成音声と，比較用の音像定位のない合成音声を提示した．被験者は何回でも繰り返し音声を聞くことができる．被験者には音声をすべて聞き終えたあとに，提案システムの音像定位感と使用感について 3 項目，音像定位のないシステムとの比較として 2 項目の，計 5 項目について 5 段階で主観評価を行ってもらった．なお，「システムの利便性」については，視覚障害者の人が利用したことを想定して，評価を行ってもらった．

1. 音像定位ありの音声の左右方向の定位感（1．とても悪い〜5．とても良い）
2. 音像定位ありの音声の前方向からの定位感（1．とても悪い〜5．とても良い）
3. 音像定位ありの音声の移動感の滑らかさ（1．とても不自然〜5．とても自然）
4. 音声としての聞き取りやすさ（1．とても悪い〜5．とても良い）
   (a) 音像定位ありの音声
   (b) 音像定位なしの音声
5. システムの利便性（1．なし〜5．あり）
   (a) 音像定位ありの音声
   (b) 音像定位なしの音声

3.3 実験結果

主観評価の結果を Table 1 に示す．左右方向の定位感は，5 つの項目で最も高い評価かつ分散の少ないことから，非常に良好で個人差も非常に少ないことが確認された．文献 [5] によると実測した両耳間時間差，両耳間レベル差を模擬することのみで左右の定位感が再現できることを示している．今回用いた文献 [4] の HRTF の補間手法も，両耳間時間差と両耳間レベル差の補間を行っていると考えることができ，そのため左右の定位感は良好であったと考えられる．

一方で，前方向からの定位感は，5 項目で最も低く，また最も分散が大きい結果となった．これは，前方向の定位感に人によるばらつきがあり，かつ全体的に定位が上手く行われないことを意味している．

Table 1　主観評価結果

評価項目		平均	分散
1. 左右方向の定位感		4.5	0.25
2. 前方向からの定位感		2.9	1.7
3. 移動感の滑らかさ		3.8	0.62
4. 聞き取りやすさ	音像定位あり	3.7	0.53
	音像定位なし	3.8	0.36
5. 利便性	音像定位あり	4.3	0.53
	音像定位なし	3.7	0.89

読み上げ音声の移動感の滑らかさは，平均 3.8 となった．これは，若干移動感に不自然さが残っていることを意味している．今回の HRTF の補間手法を提案している西野らの文献 [4] では，30° 間隔で測定した HRTF から 5° 間隔の HRTF を補間によって作成し，それによる音像の移動感を評価している．その評価結果は 5 段階（0. 悪い〜4. 良い）で約 2.5 ポイントとなっており，本研究における移動感の評価（1〜5 の 5 段階で 3.8 ポイント）もほぼ同等の結果と言える．文献 [4] では，この補間 HRTF と，5° 間隔で実測した HRTF との比較評価も行っており，後者の方が約 0.6 ポイント高い評価を得ていることから，補間アルゴリズムを改良することによって，本システムにおける音源の移動感も改善するものと考えられる．

聞き取りやすさについては音像定位ありの音声と音像定位なしの音声でほぼ同じ評価が得られた．音像定位がある音声でも，音声の聞き取りやすさには影響がないことが確認された．

システムの利便性については，音像定位ありのシステムの方が，音像定位なしのシステムに比べ高い評価を得ることができた．この項目に対して有意水準 5%で片側検定を行ったところ両者に有意差が認められた．

4 おわりに

本研究では，Web ブラウザ上に表示された文章を，その文字位置に対応する方向から聞こえるように立体的に音像定位させて読み上げを行う Web ページ読み上げシステムの提案を行った．被験者実験の結果，音像定位のあるシステムと音像定位のないシステムの両者で生成された音声の聞き取りやすさに差はほとんどなく，利便性については，音像定位なしのシステムより音像定位ありのシステムの方が高い評価を得た．

今後の課題としては，前方向の音像定位感の改善と仰角方向の音像定位を実装した場合の定位感，利便性についての検討，視覚障害者を対象としたシステムの評価などが挙げられる．

参考文献

[1] http://www-06.ibm.com/jp/accessibility/solution_offerings/hpr/hpr-view.html
[2] 山田 他，音講論 (秋)，1-2-8，pp.221-222, 2001.
[3] 外川 他，音講論 (秋)，3-10-9，pp.345-346, 2002.
[4] 西野 他，音響誌，vol.55，pp.91-99，1999.
[5] 伊藤 他，信学技報，vol.101，pp.17-24，2001.