2004－DD－45　(9)
2004／7／30

ビデオ会議における発言表示手法の提案

宮崎　観世†　瀬川　典久†　阿部　芳彦†　村山　優子†

† 岩手県立大学大学院ソフトウェア情報学研究科　〒020–0193 岩手県岩手郡滝沢村滝沢字巣子 152–52
E-mail: †g231c034@edu.soft.iwate-pu.ac.jp

あらまし　近年，ネットワークの広帯域化が進み，また，高性能な情報端末が普及し，マルチメディア情報基盤が現実的なものとなってきた．本研究では，そのような環境で映像によるコミュニケーション支援のための新しい機能を紹介する．従来の遠隔会議等の実時間型映像配信システムでは，映像に字幕を表示することにより「何が話されたか」を観衆が確認できるが，「誰が話したか」が不明瞭である．本論文では，映像に吹き出しを表示するという手法を提案し，吹き出しを表示するシステム Comicar について述べる．

キーワード　吹き出し，拡張現実感，複合現実感，音声認識

A Proposal for the Method of Displaying Utterance on Video Conferencing

Mitsugu MIYAZAKI†, Norihisa SEGAWA†, Yoshihiko ABE†, and Yuko MURAYAMA†

† Graduate School of Software and Information Science, Iwate Prefectural University
Sugo 152–52, Takizawa-mura, Iwate, 020–0193 Japan
E-mail: †g231c034@edu.soft.iwate-pu.ac.jp

Abstract　Nowadays the broadband networks have spread across, and the high performance computers are popular, so that the multimedia information infrastructure is available. In this research, we introduce a new function with video image to support communication. With the conventional real-time video conferencing system, the captions on a video image helps viewers to understand "what is said," but not to see "who says it." For example, if a crowd of people spoke simultaneously, it would be not clear to the viewers who said what. In this paper, we introduce the method of displaying a baloon-caption at a speaker on a video image, and Comicar: a system that displays the balloon-caption.

Key words　Balloon-caption, augmented reality, mixed reality, speech recognition

1. はじめに

近年，アメリカでの同時多発テロ事件やアジアでの SARS 感染などにより，現地へ赴くことに生命や健康への脅威がつきまとう状況となっている．このような状況の中で，遠隔地の人とコミュニケーションを図るための方法として，ビデオ会議システムの利用が考えられる．

現在では CU-SeeMe [1] や NetMeeting [2] をはじめ，様々なビデオ会議システムが存在している．そして，近年の情報端末の高性能化や広帯域ネットワークの普及に伴い，これらビデオ会議システムの通信品質も向上している．しかし，通信経路の混雑などの原因によりネットワークの性能を充分に得られない場合がある．このようなときには音声が途切れてしまい，発言の意味を理解することができなくなってしまう．音声品質を確保できない場合には，発表者がゆっくり話すことで解決することもある．

音声情報を字幕やテロップのような文字情報として付属することは，発言内容に対する理解を深める手段として有効である．これはテレビや映画で私たちも経験的に実感できることである．本論文では，失われた音声情報を補完するための方法としても文字情報を添付することが有効であると考え，さらにこの文字情報を吹き出しとして表現することを提案する．吹き出しとは，漫画で登場人物が喋る台詞部分を示すために使われる枠のことである．この提案手法を実装したシステムを漫画の Comic と拡張現実感の AR を組み合せて，Comicar と名付けた．

次節以降，提案手法と実装したシステムについての説明を行う．第 2 節では吹き出しの利用に関する説明，第 3 節では本システムの設計，第 4 節では実装，第 5 節では評価，第 6 節で

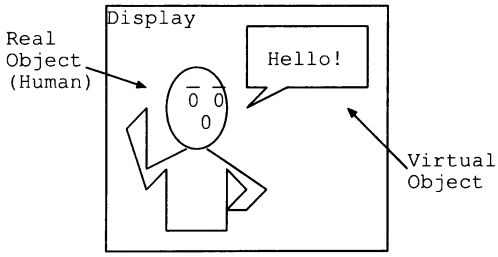


図 1　動作イメージ
Fig. 1　Operation image

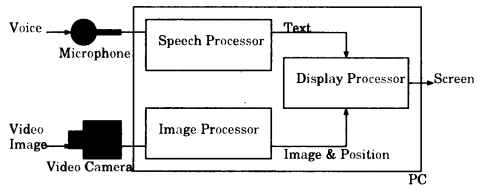


図 2　システムの構成概念
Fig. 2　Construction

は関連研究，そして第 7 節においてまとめと今後の課題について述べる．

2. 吹き出しの表示手法について

吹き出しを表示するためには，発言内容である文章と表示させる位置が必要となる．表示される文章は発表者の音声であり，表示させる位置は発表者の立ち位置である．発表者の音声を取得し文章へ変換するためにリアルタイムの音声認識技術を利用し，発表者の立ち位置を取得するためにマーカー型の拡張現実感技術を利用する．

拡張現実感技術は，現実環境と仮想環境の位置合わせの手法において，センサー型とマーカー型に分類される．センサー型とは，磁気センサーや超音波センサー，ジャイロセンサーなど，特殊なハードウェアを用いる手法である [10]．マーカー型とは，ビデオカメラからの映像を解析して現実環境の座標系を得る手法である．センサー型はハードウェア的に高コストであったり，計測可能な空間が固定されるなどの問題があるため，本システムはマーカー型の手法を用いる．

また，ここで言うマーカーとは，位置を感知するためのオブジェクトのことである．通常，特殊な色や形状の物体をマーカーとして利用する．顔の位置をパターンマッチングなどの映像処理で特定する場合には，その顔をマーカーとして利用していることになる．

図 1 は提案手法を実装したシステムの動作イメージである．Real Object とは人や周囲の環境のことであり，現実に存在する物である．Virtual Object とは吹き出しのことであり，CG で描画された仮想の物である．

発言を吹き出しとして表示することによって，次のような利点がある．

- 発表者の表情などのノンバーバルな情報も合わせて取得できる．
- 耳で聞き逃した内容を目で見て補完できる．
- 誰が喋ったかが一目で理解できる．

発言の音声認識結果が正確ではなかった場合において，発表者の表情や身振り手振りといったノンバーバルな情報によって観衆が本来の意味を類推することで，正しい発言内容を理解することができるという効果がある [7]．

吹き出しを表示する際には，表示画面領域内に吹き出し全体を収めなければならない．吹き出しは発表者を識別するマーカーに追従して移動する．そのため，画面端にマーカーがある場合を考慮して、映像内のマーカーの位置により、吹き出しの相対的な表示位置を変更する．吹き出しはその内部の文章の長さに応じて大きさが変化する．そのため，文章の長さによる表示位置の調整も必要となる．

縦書きや横書きといった文章の並び方はその文章の読みやすさに関係する [7]．しかし，読みやすい文章の並び方は，生活習慣や好き嫌いに影響され人により異なる．TV の字幕やテロップなど，横書き文章への慣れによるためか，縦書き文章はビデオ会議システムへの表示には向いていないようである．そのため本システムでは，吹き出し内の文章は横書きで表示する．更に，横書き文章の表示方法も人により読みやすさは異なる．利用者が選択できるように複数の表示方法を用意することが望ましい．

発表者の音声の音量により，吹き出し内の文章の大きさを変える処理を行う．発表者が力強く大きな声で発言しているにも関わらず，吹き出し内の文字を小さく表示してしまうと，発表者の実際の様子と異なる．発表者がどのような調子で発言しているかを，発表者の映像だけでなく文字の大きさでも表すことで，発言に対する観衆の理解が深まると考える．

また，発言の音声認識処理中であるという状態を発表者や観衆に対して“…”などの表現で，吹き出し内の文章の一部として提供することも重要である．これにより，発表者は自分の発言が入力されたかを確認でき，観衆は“文章が表示される”という予測を行い，吹き出し内の文章を読む姿勢を取ることができる．

3. システムの設計

本システムは，音声処理部，映像処理部，表示処理部の三つで構成されている．図 2 は本システムの構成概念図である．

音声処理部 (Speech Processor) では，マイクロフォンから入力された発表者の音声を音声認識技術を用いて文章へと変換する処理を行う．まず，音声処理部は音声認識エンジンに音声の処理を依頼するための設定と接続を行う．次に，マイクロフォンへ入力された発表者の音声は音声認識エンジンによって文章へと変換される．そして，音声処理部は音声認識エンジンより送られるメッセージ内から変換された文章を取りだして保存し，吹き出し作成に利用する．

映像処理部 (Image Processor) では，ビデオカメラで撮影さ

表 1　マイクロフォンの性能
Table 1　Microphone specification

製造社	PLANTRONICS
製品名	.Audio70
感度	-39dBV/Pa +/-5dB
タイプ	エレクトレット，単一指向性
周波数特性	100Hz - 8kHz
インピーダンス	~3k Ω

表 2　ビデオカメラの性能
Table 2　Video camera specification

製造社	SONY
製品名	DCR-VX2000
CCD 総画素数	1/3 型 38 万画素× 3
CCD 有効画素数	34 万画素× 3

れた映像からマーカー型の拡張現実感技術を用いて発表者の位置を取り出す処理を行う．まず，本システムで使用されるマーカーのパターンデータを読み込み，ビデオカメラの設定を行う．次に，映像処理部はマーカーを付けた発表者が撮影された映像をビデオカメラから受け取り，マーカー検索処理を行う．そして，本システムが使用するパターンが存在するかを判断し，存在するならば，マーカーの位置を取りだして吹き出し作成に利用する．

表示処理部 (Display Processor) では，音声処理結果の文章と映像処理結果の位置を利用して，吹き出しを描画する処理が行われる．まず，表示処理部は吹き出しを描画するための設定を行う．次に，音声処理部と映像処理部から送られてきた文章と位置情報を利用し，吹き出しの作成を行う．そして，ビデオ映像の上に重ねて吹き出しの描画を行う．マイクロフォンへの音声入力がない場合や，マーカーが隠れていて発見できない場合には，ビデオカメラの映像だけを表示する処理が行われ，吹き出しは描画されない．

4.　システムの実装

ハードウェアの構成，ソフトウェアの構成と開発環境を説明する．そして，実装されたシステムの表示画面の様子を説明する．

本システムで使用したハードウェアは，マイクロフォン，ビデオカメラ，PC である．各々の性能は，表 1，表 2，表 3 に示す．ビデオカメラと PC は IEEE1394 で接続する．マイクロフォンは PLANTRONICS 社の.Audio70 を使用した．ビデオカメラは SONY 社の DCR-VX2000 を使用した．PC は DELL 社の INSPIRON I8200 を使用した．CPU は Intel Pentium4(1.7GHz) である．映像の伝送速度は，720 × 480 の画面を TrueColor 24bit で約 27frame/sec であるので，約 27MByte/sec である．IEEE1394 の最大伝送速度が 50MByte/sec であるので，約半分の伝送速度である．

本システムで使用したソフトウェアは，音声認識として IBM 社の ViaVoice V10 と ViaVoice SDK 8.0 [8]，拡張現実感としてワシントン大学の Human Interface Technology Laboratory

表 3　PC の性能
Table 3　PC specification

製造社	DELL
製品名	INSPIRON I8200
CPU	Intel Pentium4 1.7GHz
ビデオカード	GeForce4 32MB

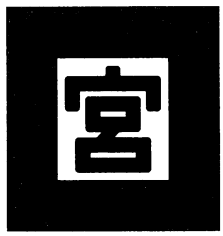


図 3　パターン例
Fig. 3　A pattern example

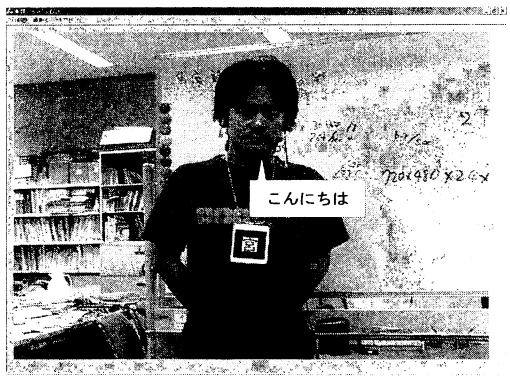


図 4　システムの表示画面
Fig. 4　The display screen of the implemented system

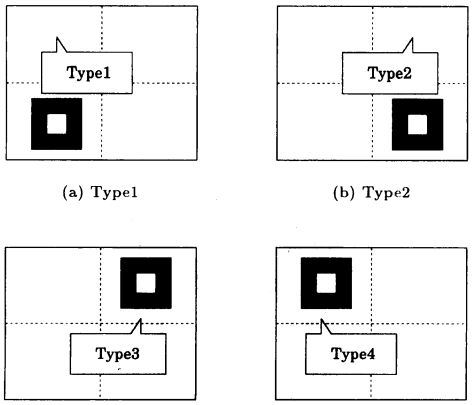
の ARToolKit [5] [6] である．

ARToolKit とは，拡張現実感アプリケーションを簡単に開発できるようにするためのライブラリとそのツールキットである．ARToolKit のライブラリは，現実環境と仮想環境の位置合わせが容易に実現でき，リアルタイムでマーカーの追跡を行う．ARToolKit は主プラットフォームである Linux の他，IRIX，Windows，Mac OS X など様々なプラットフォームで動作する．ARToolKit のライセンスは GPL となっていて，同研究所の Web サイトから各々のバージョンを入手できる．

今回の実装では，位置特定のマーカーとしてこの ARToolKit のマーカーを利用した．ARToolKit が検索できるマーカーのパターン例を図 3 に示す．本システムは，このマーカーを名札のように発表者の胸に付けることで，マーカーの位置を発表者の位置とし，吹き出し描画に利用している．システム開発には，Microsoft 社の Visual C++ 6.0 を使用して，WindowXP 上に構築した．吹き出しの描画は OpenGL を用いて行った。

図 4 は，実装したシステムの表示画面の様子である．

表示画面領域内に吹き出し全体を収めるために，映像の中心



(a) Type1　(b) Type2

(c) Type3　(d) Type4

図 5　吹き出しの表示位置

Fig. 5　Display position of the balloon-caption

を二次元座標の原点とし，マーカーの座標によって，マーカーとの相対的な吹き出しの表示位置を四つに分類した．図 5 はマーカーの座標による四種類の吹き出しの表示位置である．図 4 の吹き出しの表示位置は図 5 の (a)Type1 であることが確認できる．

吹き出し内の文章は，一行最大 16 文字まで表示するよう実装した．新しい文章は古い文章の後に追加される．16 文字を越えて表示できなかった文字は，新しい吹き出しとして表示される．少し時間が経過してから発言すると，文章は新しい吹き出しに表示される．今回，実装した吹き出し内の文章の表示方法は一つであったが，複数用意することが望ましい．

5. 評　　価

システムが表示する文章の正確性を評価するために，音声認識率を計測する．リアルタイム性を評価するために，発言が吹き出しとして表示されるまでの表示反応時間を計測する．そして，本システムのデモンストレーションを行ってアンケートを集計した．

音声認識率は，定型文とそれを読んで認識された文章を比較して測定する [4]．計算式は次の通りである．

$$R = \frac{N_t - N_e - N_i - N_d}{N_t} \times 100 \quad [\%]$$

R は音声認識率，N_t は全字数，N_e は誤変換字数，N_i は誤挿入字数，N_d は非認識字数である．使用する定型文は 3 種類の文語である．各々を表 4 に示す．読む速度は，単語ごとに区切って喋らずに，一つの文章毎に少しの間を開けるぐらいの普通の速度である．読む回数は各々10 回ずつである．

表示反応時間は，発言を開始してから吹き出しが表示されるまでの時間の平均値を求めた．読む文章により表示反応時間に差が出る可能性があるので，5 種類の文章を読む．読む回数は各々5 回ずつである．

表 4　音声認識率測定に用いた定型文

Table 4　Texts to measure the speech recognition rate

定型文 1	画面に出ている文を声に出して読んでください． ここでは，句読点などの記号を読む必要は ありません． この声を録音して，特徴を分析します． 声の特徴に合わせて学習するためです． 同時に，使用環境の傾向も調べます． 読み終わったら，分析作業が終わるまで お待ちください．
定型文 2	音声認識システムは，口述タイプライタと 呼ばれたりします． このシステムのプロトタイプ名は， 「タンゴラ」でした． この名前は，口述タイプライタに関係があります． タンゴラは，スピード持久タイプライティングの ギネスブック記録保持者です．
定型文 3	有用な人材を確保するために， 企業は SPI 検査や筆記試験など， さまざまな方法を採用する． しかし，志望者の性格や能力を見きわめる最終的な 判断は，面接試験で下される． そのうえ就職協定の廃止になって， 企業が採用活動に十分な時間をかけられるように なったため，「面接」を 非常に重視する傾向が強まっているのだ．

表 5　音声認識率の結果

Table 5　Results for the speech recognition rate

文章	平均 [%]	最低 [%]	最高 [%]
1	93.91	73.28	99.14
2	92.59	87.38	96.12
3	92.36	85.40	97.08

デモンストレーションに対するアンケートでは，文章の意味の理解と文章が表示される速度についての点数を集計する．点数は 0.0 から 5.0 までの間で採点する．本システムの全体的な感想や意見も集計し好印象を悪印象に分類する．アンケートの被験者は健康な 14 人の男女である．今回のアンケートでは，被験者が用紙に記入を行う前にこのシステムの概要について簡単な説明を行った．アンケート結果は，システムを動作させた状態で更に 10 分間程度の説明を行い，吹き出しが表示されている画面の様子についての点数である．

アンケート項目は次の通りである．

（1）　文章の意味の理解

（2）　文章が表示される速度

（3）　システムに対する感想や意見

各々の結果を表 5 から表 8 に示す．

音声認識率は高い数値を得られた．しかし，人名などの固有名詞の認識と同音異句語の変換が困難であることが比較結果から判明した．また，アンケートの被験者の感想から，明瞭に発音する必要があることが判明した．読む速度と内容理解の上で

表 6　表示反応時間の結果

Table 6　Results for the reaction time

文章	時間 [sec]
1	2.2
2	2.5
3	3.0
4	3.1
5	3.3

表 7　アンケートの点数

Table 7　Questionnaire marks – Likert scale evaluation

項目	平均	標準偏差	変動係数
1	2.379	1.101	0.463
2	2.843	0.978	0.344

表 8　アンケートの感想

Table 8　Questionnaire comments

好印象	悪印象	回答なし
7	3	4

観衆に対して認識結果全文を提示してしまうことは，吹き出し内に表示される文章に文字数の制限があるため問題となる．リアルタイムニュース字幕の内容了解度による評価の試み [9] の実験結果によれば，要約文と全文で内容理解に差は生じない．全文提示では要約処理の手間は省けるが，可読性を損なう懸念がある．要約文提示では内容を理解しやすいが，負荷が大きくなる．この問題の解決は今後の課題である．

表示反応時間は文章により差はあるが，平均して約 3 秒という結果が得られた．しかし，アンケート結果からは少し遅いと考えられる．実際の映像との同時性を高めるため，この約 3 秒という表示反応時間を短縮する改良を行う予定である．長い文章を表示する場合には，各吹き出し毎の表示時間を調整して理解しやすくする必要もある．

アンケートの結果，本システムの動作に対して良い印象を持った人と悪い印象を持った人に別れた．そして，項目 1 と 2 のどちらもそれほど高い平均点ではないので，改良を行う必要があると言える．

これらの問題点を解決していくことで，利用者に対してより理解しやすい情報の提示を行うことができる．アンケートの感想からは，システムの性能を向上させて実現できた場合に役立つという意見が多く，吹き出しを表示することは発言を理解することに良い影響を与えることが期待される．

6. 関連研究

音声認識技術での関連研究として，B.U.G. 社が開発した音声同時字幕システムがある [11]．このシステムは，国際ユニバーサルデザイン会議や東京大学の講義などにおいて実際に運用されていて，発表者の発言を訓練された同時復唱者が復唱し，その音声を音声認識技術により文章化し表示するという方法を取っている．このシステムの表示反応時間は約 3 秒であるため，本システムの表示反応時間はリアルタイム字幕システムにおいて妥当な時間と言える．

拡張現実感技術での関連研究として，Rasa という付箋システム [12] がある．このシステムは，地図などの作業領域をビデオカメラで撮影し，その領域上に付加情報をプロジェクターで表示するシステムである．領域内で何らかの役割を持つオブジェクトを付箋で表し，音声認識とジェスチャーを用いてオブジェクトの移動や選択，情報追加などの作業を行う．このシステムは，拡張現実感と音声認識を組み合わせている点で本システムと似ている．

また，会議の議事録を吹き出しを用いて HTML 文書として提供するシステム [13] がある．このシステムは，ビデオカメラで撮影された映像からダイジェスト画像を抽出し，議事録の要約文をその画像の代替文のように吹き出しとして表示する．情報提示に吹き出しを利用するという点で本システムと同じであるが，“誰の発言か” という情報は状況的に判断できるが明確ではない点において異なる．

チャットシステムにおいて，吹き出しを発言表示に利用することはよくある．Comic Chat [14] というシステムは，チャット上でのコミュニケーションを漫画の 1 コマのように限られた枠の中で表現している．吹き出しの表示位置や大きさ，吹き出しの尻尾の表示位置を決定するための手法などに関して，今後，本システムの吹き出し表示指針を見直す場合に非常に参考となるシステムである．

7. ま と め

吹き出しを用いた発言の表示手法を提案し，音声認識技術と拡張現実感技術を用いて吹き出しを表示するシステム Comicar について述べた．吹き出しを表示することの利点や表示に関する指針について述べ，システム設計を行った．そして，ViaVoice, ARToolKit, OpenGL を用いて本システムを実装した．本システムの有用性を評価するために，音声認識率と表示反応時間を計測し，デモンストレーションに対するアンケートを集計した．その結果，本システムに対する評価は様々であり，現段階では良くも悪くもないと言える．本システムの更なる改良が必要であり，それにより有用性が向上する可能性がある．

今後の課題は，今回のアンケートの意見を反映した本システムの改良や，音声なしの吹き出し表示のみでの本システムの有用性を調査することである．また，本システムをビデオ会議システムとして実際に使用するために，発表者が多人数の場合やネットワークを利用して遠隔地とのコミュニケーションを行う場合などを想定した機能追加を行う．更に，本システムの音声認識結果を翻訳ソフトによって他言語に変換することで，異国間でのコミュニケーションに応用が可能である．

文　献

[1] T. Dorcey: “CU-SeeMe Desktop Video Conferencing Software”, Connexions 9, Vol.3, 1995.

[2] Microsoft corporation: “NetMeeting Home”, http://www.microsoft.com/windows/netmeeting/

[3] National Captioning Institute http://www.ncicap.org

[4] ANDO,IMAI et al.: “Simultaneous Subtitling System for Broadcast News Programs with a Speech Recog-

nizer(Special Issue on the 2001 IEICE Excellent Paper Award)", IEICE Transactions on Information and Systems, Vol.E86–D, Num.1, pp.15–25, 2003.

[5] M. Billinghurst, H. Kato, E. Kraus, R. May, Shared Space: Collaborative Augumented Reality. In Visual Proceedings, SIGGRAPH 99, August 7-12th, Los Angeles, CA, ACM Press 1999.

[6] H. Kato, M. Billinghurst: "Marker Tracking and HMD Calibration for a Video-based Augmented Reality Conferencing System", IWAR'9, pp.85-94, 1999.

[7] 小坂井, 奈良他: "音声認識技術と透過型 HMD を利用した聴覚補助方式", 電子情報通信学会技術研究報告, Vol.100, Num.712, pp.35-41, 2001.

[8] IBM Japan, Ltd.: "IBM ボイスらんど" http://www-6.ibm.com/jp/voiceland/

[9] 小峰, 星野他: "リアルタイムニュース字幕の内容了解度による評価の試み", 電子情報通信学会技術研究報告, Vol.97, Num.529, pp.73-78, 1998.

[10] 藤井 博文, 神原 誠之, 岩佐 英彦, 竹村 治雄, 横矢 直和: "ジャイロセンサを用いたビジョンベースド AR のためのマーカ追跡手法", 電子情報通信学会技術研究報告, Vol.99, Num.488, pp.31-36, 1999.

[11] 株式会社ビー・ユー・ジー: "音声同時字幕システム" http://www.bug.co.jp

[12] David R. McGee, Philip R. Cohen, Lizhong Wu: "Something from nothing: augmenting a paper-based work practice via multimodal interaction", Proceedings of DARE 2000 on Designing augmented reality environments, pp.71-80, 2000.

[13] Patrick Chiu, John Boreczky, Andreas Girgensohn, Don Kimber: "LiteMinutes: an Internet-based system for multimedia meeting minutes", Proceedings of the tenth international conference on World Wide Web, 2001.

[14] David Kurlander, Tim Skelly, David Salesin: "Comic Chat", Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, pp.225-236, ACM Press 1996.