2009－SLP－75 (14)
2009／2／7

会議の情報保障を目的とした吹き出し型字幕提示方式の検討

藤井 絢子　　南條 浩輝　　吉見 毅彦

龍谷大学 理工学部 情報メディア学科
〒 520-2194 大津市瀬田大江町横谷 1-5
e-mail: {fujii,nanjo,yoshimi}@nlp.i.ryukoku.ac.jp

あらまし　近年，教育機関をはじめとする様々な場面において障がい者に対する情報保障の充実が求められている．情報保障を行うためには人材の確保・育成が必要であり，膨大な労力を要する．音声認識技術は聴覚障がい者への情報保障に有望な技術であり，実際に多くの研究が盛んに行われている．それらの多くは放送ニュースや講義，講演，もしくは国会などに代表されるよくコントロールされた会議の字幕付与の研究である．これらではいかに音声認識の精度を高めるかが主な研究ターゲットであり，どのように字幕を提示するかについてあまり注意を払ってこなかった．実際に，これらでは話者は基本的に 1 名であるため，字幕提示の方式を考慮する必要はそれほど高くない．一方，複数の話者が存在する会議での情報保障の研究は十分でない．本論文では，複数人での話し合いにおける会話内容を，吹き出し型で表示する方式について検討を行ったので，その結果を報告する．

キーワード　情報保障，会議，吹き出し，字幕付与，音声認識

Speech Baloon Captioning System for Information Support on Meetings

Ayako Fujii　　Hiroaki Nanjo　　Takehiko Yoshimi

Department of Media Informatics, Faculty of Science and Technology,
Ryukoku University
1-5 Yokotani, Oe-cho, Seta, Otsu, 520–2194

Abstract　Information support to handicapped people is addressed. Automatic speech recognition which converts speech to texts is promising for hearing support, and several studies have investigated; such as automatic captioning for TV program, automatic transcription of oral presentations, lectures, and meetings. These conventional studies mainly focused on how to recognize speech accurately, and did not pay attention to how to display caption texts. Actually, in TV news, oral presentations, and lectures, a single speaker usually talks, and thus, how to display caption texts has not been a significant problem. In this paper, we focus on the information support on meeting which includes several speakers, and describe a novel captioning system which displays captions with speech balloon based on automatic face detection and speech recognition.

Key words　Imformation support, Meeting, Speech Baloon Captioning, Automatic Speech recognition

1　はじめに

近年，教育機関をはじめとする様々な場面において障がい者に対する情報保障の充実が求められている．聴覚障がい者への支援では，聴覚的情報を視覚的情報に変えて伝達する手段が採られており，手話やノートテイクなどがその典型例として挙げられる．手話通訳は即時性が高く，ディスカッションが中心のゼミや実習などで活用されている [1]．しかし，手話通訳の能力を必要とするため，支援のコストが大きい．ノートテイクは特別な準備を必要としないものの書くスピードは話すスピードよりも遅く，遅れが生じてしまう．この遅れを埋めるためにある程度要約して書き進める必要があり，発話内容が 100%伝達されないという問題もある．このように，現状では聴覚的なハンディキャップを持つ人への支援はまだまだ不十分である．

これまで，このような情報保障は人手作業で行われており，膨大な時間と労力を必要としていた．音声を文字に変換する音声認識技術は，聴覚障がい者への支援にとって有望であり，音声認識技術を用いた情報保障の取り組みが行われている [2][3][4][5]．また,『聴覚障害者のための字幕付与シンポジウム』が京都大学で開催されるなどその注目も高い．本研究も，このような音声認識を用いた字幕付与システムに焦点をあてている．

音声の情報保障を行う際には，どのように字幕を表示するかも検討すべき課題である [3]．TV や映画の字幕表示では，発言者の発話が画面の下もしくは横に表示されているため発言者と字幕文字が近く，表情やしぐさなどの非言語情報と言語情報との統合が比較的容易と考えられる．一方，講演や会議では，専用のスクリーンに文字が表示されることが多く，情報保障を受ける側は，話者と字幕文字を交互に見る必要がある．発言者の映像と発言内容がひとつの画面に表示される方が発話者が発信している情報を受け取りやすく，保障を受ける側にとっては情報保障の観点から望ましいと考えられる．また，唇の動きから発話文を読み取る（リップリーディング）ことができる人にとっては，字幕が顔付近にある方がリップリーディングの結果と字幕の両方を読み取りやすい可能性も考えられる．本研究でも，このような映像と字幕をひとつの画面に表示するタイプの情報保障を考える．講演や講義などでは話者は基本的に 1 人であり，この場合は映像に TV や映画のように字幕を表示することで十分な情報保障ができると考えられる．一方，会議などの話者が複数人存在する場合は，そのような表示方式では発言者の特定が困難であり，適切に情報保障を行えない可能性がある．この問題に対し，本研究では字幕をそれぞれの話者の顔付近に表示する吹き出し型の字幕表示方式を提案する．吹き出し型での字幕表示はこれまでにも見られるが [6][7]，いずれも情報保障を目的としたものではなく，この点において本研究とは異なる．

2　吹き出し型字幕提示方式の評価

2.1　調査の概要

話者が複数人存在する会議の情報保障という観点で，吹き出し型字幕提示方式の有効性の調査を行う．今回は，健聴者を被験者として，吹き出し型字幕提示方式の有効性を調査を行う．

また，議長がいる会議では，次に誰が発言するかを比較的予測しやすいと考えられる．一方，議長がいない会議では，誰がいつ発言するかの予測が難しく，発話のタイミングが重複することも十分に考えられる．そこで，議長の有無で字幕の表示形式に差があるのかについても調査を行う．

2.2　評価方法

評価はアンケート調査に基づいて行った．具体的には，TV や映画の字幕のように画面の下部にまとめて字幕を提示する方式（以下，TV 字幕型表示方式）と提案する顔付近に字幕を表示する方式（以下，吹き出し型字幕表示方式）の映像を提示し，各映像を見終った後にアンケートに回答してもらった．映像に音声は含まれていない．会議形態ごとの字幕の表示方式の有効性を調査するため，会議の形態を議長がいる場合といない場合に分け，それぞれについて TV 字幕型表示方式と吹き出し型表示方式の映像（合計 4 種類）を用意した．提示したデモ映像の種類とそのイメージを図 1 に示す．

被験者にはこの 4 つの映像を順に見てもらい，各映像を見るごとにアンケートに回答してもらった．映像を提示する順序の評価への影響を除くために，

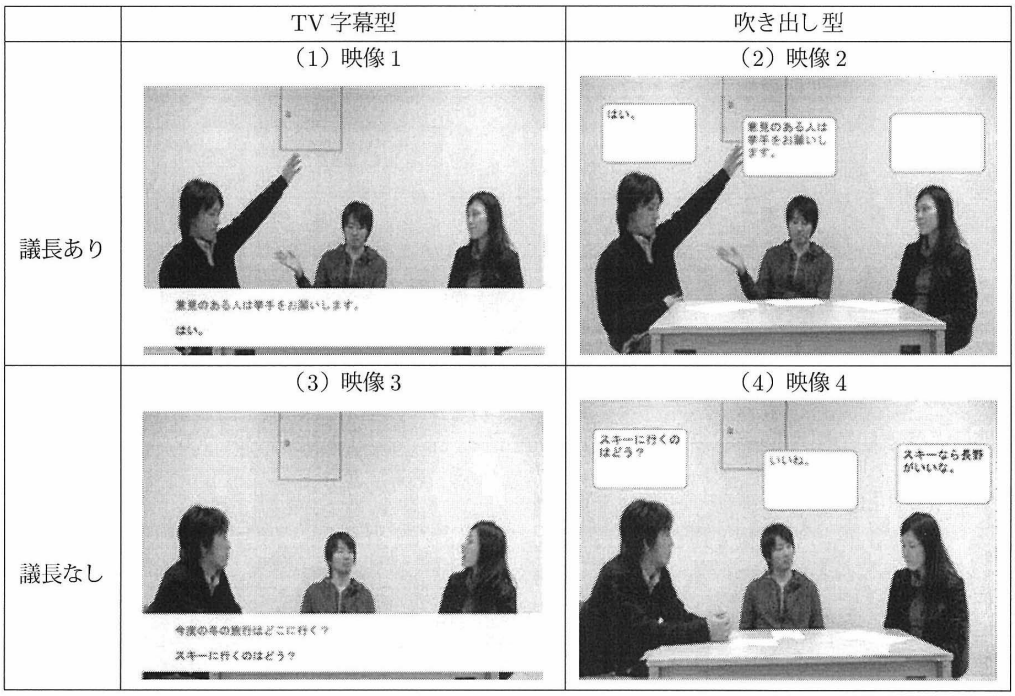

図 1: デモ映像の種類とそのイメージ

表 1: デモ映像の試聴順序

パターン	試聴順序
A	映像 1 → 2 → 3 → 4
B	映像 2 → 1 → 4 → 3
C	映像 3 → 4 → 1 → 2
D	映像 4 → 3 → 2 → 1

表 1 に示す 4 パターンの提示順序を用意し，被験者ごとにパターンを変えて提示した．なお，映像の内容は，3 名の会議参加者が旅行の行き先を決めるというものである．議長ありの映像では，中央の人物が議長となって参加者の発話をコントロールしており，発話に重複はない．議長なしの場合は，3 名の参加者がそれぞれのタイミングで発話を行っている．映像の長さは全て 1 分程度である．今回は字幕提示方式自体の評価が目的であるため，字幕には音声認識誤りは含まれていない．

被験者には，10 項目の質問それぞれに 7 段階評価を行ってもらった．質問項目を図 2 に示す．また，アンケートの最後に自由記述欄を設け，被験者に意見を求めた．

質問 1：字幕位置は適切であった
質問 2：字幕は読みやすかった
質問 3：会議の話の流れが理解できた
質問 4：臨場感があった
質問 5：実際にこのシステムを使って会議に参加できると思う
質問 6：楽しい
質問 7：使いたい
質問 8：心強い
質問 9：親しみやすい
質問 10：役立つ

図 2: 質問項目一覧

2.3 調査結果

2.3.1 分析方法

アンケートの各項目の評点に対して 2 要因の分散分析 [8] を行った．各要因の主効果が見られるかを 5%水準で検定した．

2.3.2 分析結果

アンケート調査の被検者は龍谷大学理工学部情報メディア学科の学生 21 名である．各項目の評点の平均値を図 3 に示す．吹き出し型の提示を行った場合は，評点平均 5 以上のものが多く，TV 字幕提示に比べて高い評点が得られていることがわかる．

分散分析を行ったところ，議長あり／なしという会議形態の主効果が 10 項目中 1 項目（「質問 8：心強い」）で見られた．次に，TV 字幕型か吹き出し型かという字幕の表示形式の要因では，10 項目中 8 項目で主効果が見られた．主効果がみられた評価項目を図 4 に示す．

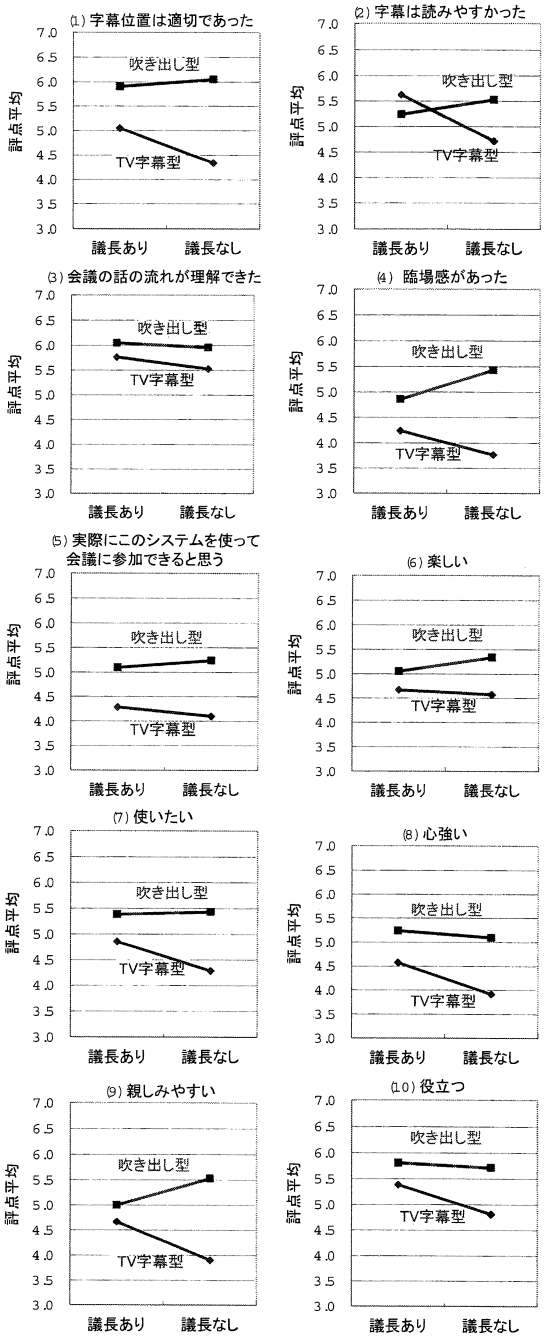

図 3: 各項目の評点の平均値

また，図 5 に示す 3 つの評価項目では交互作用の有意差が確認できたため，Bonferroni 法を用いて単純主効果の検定を行った．具体的には，調整化された有意水準 $\alpha\prime$ を求め，各比較ペアの t 検定の結果によって算出された確率値に対して判定を行った．今回，有意水準 0.05 を比較ペア数 4 で割った $\alpha\prime$=0.0125 を基準値とした．

「質問 2：字幕の読みやすさ」では TV 字幕型表示方式のときの議長あり／なしにのみ有意差が見られた.「質問 4：臨場感」および「質問 9：親しみやすい」では，議長がいない場合の TV 字幕型表示方式と吹き出し型表示方式に有意差が見られた．どちらも字幕の表示形式の要因に主効果が見られているので，この結果は議長がいる場合は字幕提示方式に差がないが議長がいない場合に吹き出し型が有効であることを示している．

これらの結果をまとめると次のようになる．すなわち，議長あり会議の場合は，評価項目 1，5，6，7，8，10 で吹き出し型が有効，議長なし会議の場合は，評価項目 1，4，5，6，7，8，9，10 で吹き出し型が有効,「項目 2：字幕は読みやすかった」および「項目 3：会議の話の流れが理解できた」には差がないことがわかった．吹き出し型は使用感

質問 1：字幕位置は適切であった
質問 4：臨場感があった
質問 5：実際にこのシステムを使って会議に参加できると思う
質問 6：楽しい
質問 7：使いたい
質問 8：心強い
質問 9：親しみやすい
質問 10：役立つ

図 4: 字幕の提示方式に主効果がみられた評価項目

質問 2：字幕は読みやすかった
質問 4：臨場感があった
質問 9：親しみやすい

図 5: 交互作用がみられた評価項目

においてユーザをより満足させる提示方式といえる．さらに議長なしの場合は，臨場感と親しみも得られるということが読み取れる．これらのことは，吹き出し型提示方式は，言語情報の伝達は阻害せず，非言語情報をより多く伝える方式であることを示唆している．

自由記述であげられた意見を図 9 に示す．発話者の口の動きと字幕が表示されるタイミングとの時間差が違和感を感じさせているのではないかと推測できる．また，吹き出し型字幕表示では一度に表示できる文字数が TV 字幕型よりも少なく，読みやすさの評点は平均としては差がみられなかったが，読みづらさを感じるユーザがいたこともわかる．発話文が長い場合にどのように表示するのかを検討する必要がある．

3　吹き出し型字幕提示システム

2 章で述べたアンケート評価結果より，吹き出し型での字幕提示方式が有効であることがわかった．このことに基づき，吹き出し型字幕提示システムを試作した．本章では，そのシステムについて述べる．

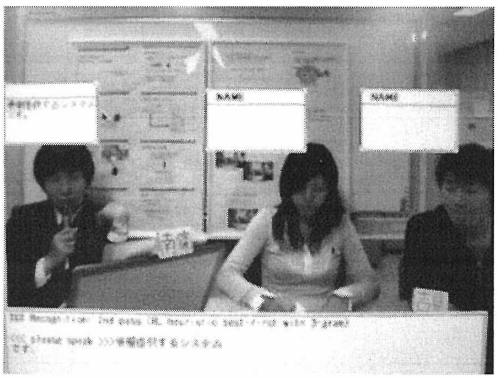

図 6: システム実行画面

3.1　システムの概要

大学のゼミなど小規模の会議を想定し，特別な機材や設備を必要とせず，持ち運び可能なシステムを設計した．具体的には，情報保障を受ける聴覚障がい者にノート PC を渡し，そのディスプレイを見ながら会議に参加できるようなシステムを設計した．このノート PC のディスプレイには Web カメラで撮影された会議の様子と会話音声の字幕が吹き出し型の字幕形式で表示される．情報保障を受ける側は，自分の意志で Web カメラを動かし，自由に視点を変えることが可能である．ノート PC 上に表示されるシステムの実行画面を図 6 に示す．

3.2　システムの構成

システムの構成を図 7 に示す．本システムは，音声部分の処理と映像部分の処理の 2 つからなる．具体的な処理の流れを以下に述べる．まず，Web カメラを用いて映像撮影を行い，同時にマイクを用いて音声の収録を行う．次に，撮影されている映像に，音声認識結果を吹き出し型画像にしたものをオーバーレイさせる．これらの処理を繰り返し行う．

次にそれぞれの処理について述べる．

3.2.1　音声認識部

音声認識部は，Julius[9][10] と一般的な HMM 音響モデルおよび N-gram 言語モデルを用いて構成した．なおこれらのモデルは，会議の参加者や内

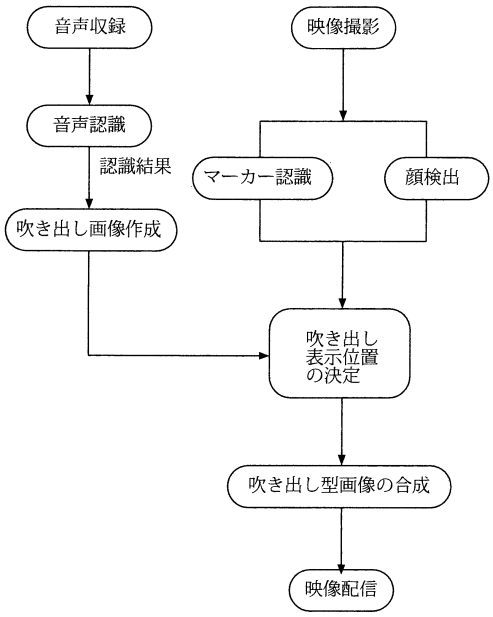

図 7: システム構成図

容によってその都度適切に選択すればよい．音声認識部では認識結果を吹き出し用の画像に出力するところまでを行う．具体的には，Julius が出力する認識結果の発話文のみを抽出し，発話者 ID と抽出した発話文を吹き出し型のテンプレート画像に埋め込み，吹き出し型字幕画像を作成する．この時，発話文の文字数を取得し，1 行につき 6 文字単位で最大 3 行までを 1 つの吹き出し型画像に埋め込む．発話文の文字数がこれを超える場合は，新しい吹き出し型字幕画像を作り，そこに文字列を埋め込む．これを全ての発話文を表示できるまで繰り返し行う．また，新しい音声の入力があれば，その都度，吹き出し型字幕画像を作成する．なお，小人数の会議を想定しているため，各参加者にはヘッドセットマイクをつけてもらうことを想定しており，話者はマイクチャンネルの ID で同定できる．

現状では，吹き出し型字幕画像が表示された後，その画像は新しい画像に更新されてしまうため，さかのぼって字幕を表示させることができない．今後，発言内容をさかのぼって閲覧できるように改良したいと考えている．アンケートの自由記述の意見を活かして，吹き出し型字幕のスクロールなども可能にする予定である．

図 8: 顔検出の様子

3.2.2 画像処理部

映像部分の処理は OpenCV[11] を用いて行った．まず，撮影された映像から顔を検出する．顔検出の様子を図 8 に示す．今回のシステムでは，発話者の位置の入れ替わりや，大きな移動がないことを条件としており，x 座標を基準として話者を特定する方式を採用した．吹き出し字幕画像を出力する位置は，検出した顔の大きさや画像全体の大きさを考慮した上で，顔の中心点から左上方向の位置に表示することとした．音声部分で作成された吹き出し型字幕画像を読込み，撮影した映像にオーバーレイして，映像を生成する．

現状では，顔の検出が不十分なため，物体追跡や肌色追跡をする必要があると考えている．また，カメラを自由に動かした時に，人物特定をするためにネームプレートなどのマーカーを検出できるようにする予定である．

4 おわりに

会議の情報保障の方式として，吹き出し型字幕提示方式を提案し，その有効性を確認した．提案する吹き出し型字幕提示方式は使用感においてユーザをより満足させる提示方式であり，TV 字幕型提示方式に比べて，非言語情報をより多く伝えられる可能性があることを示した．今後は，実際に情報保障を受ける人を対象とした調査が必要である．また，実際に会話音声の認識を行い，その認識結果を吹き出し型で表示するシステムを試作した．

・音声認識してから字幕を表示しているから仕方ないのかもしれないが，映像で映っている人の口と字幕のタイミングが合っていないのに少し違和感を感じました．
・文字が表示されるまでと口の動きからいい終わってからのラグが長い部分があったので，口が動いている間もしくはいい終わったらすぐに表示されると臨場感が増し，前半のように文字がまとめて下に出てもわかると思う．
・口の動きと文字がずれてて（タイムラグ）気持ち悪い．
・動画はリアルタイムで流すのではなく，字幕の出るタイミングに合わせて遅らせたらどうかなと思った．
・実際に映像の中の人物が口を動かしている瞬間と字幕が出るタイミングは同じなのか?
・自分の見た感じでは，バラバラ（ある程度いっしょかもしれないが）のように感じた．
・発言は一気に表示するのではなくて随時表示した方が分かりやすい．
・人の上に字幕が出るとき，瞬間的に消えるのではなく，徐々に消えて行く方が良かった．
・もう少し文字の出るタイミングを早くできれば，より臨場感が上がると思う．
・個別に吹き出しで言葉を発するときも，コメントログを残したほうがより使いやすい気がします．
・会話のログも多少欲しい．
・長い文章の場合，字幕をスクロールして出してみてもいいかと思う．
・字幕は話者の上にあった方が読みやすかった．
・話者の上に吹き出しがある方が誰が何を話しているのかわかりやすい．
・全て見終った後，考えてみると意外に個人個人の近くに字幕がついている方が理解しやすいように思いました．
・映像1と3より映像2と4の方が見やすかった．
・映像1と3の下に字幕が出る映像は，途中で誰が話しているのかわからなくなった．
・人の上に字幕があった方がわかりやすいけど，枠が小さくなっていたので文字が少し読みにくかったです．
・文字の読みやすさでは，下に大きく窓が出ている方がいいと思う．下に大きく窓を出す場合は，誰が話しているかもう少しわかりやすくする必要があると感じた．
・1〜4作品の中で，映像3が一番見やすかったです．
・文字が画面の下にある方が見やすく分かりやすかったが，文字の色だけでは誰の発言か分からなくなることがあったので，文字の横に名前があればより良かったと思いました．
・下に全員のコメントがまとめて表示する際，どの色が誰を示しているかが表示されないのはよいのかな?と感じました．(最初少し戸惑いました)
・耳が聞こえない人にはとても優しいシステムだと思います．聞こえる人も聞きづらかったことが目で見てわかりやすいので素敵です．

図 9: 自由記述で書かれた意見（一部）

参考文献

[1] 吉川あゆみ, 太田晴康, 中島 亜紀子 (他). 大学ノートテイク支援ハンドブック. ISBN-978-4-931388-52-9. 株式会社人間社, 2007.

[2] 南條浩輝. 多言語音声の同時認識枠組みの提案. 情報処理学会論文誌, Vol. 49, No. 12, pp. 4044–4048, 2008.

[3] 水島昌英, 織田修平, 政瀧浩和, 古家賢一, 片岡章俊. 音声認識による会議支援情報保障システム使用時の話者及び訂正者の負担度の評価. 電子情報通信学会技術研究報告, TL2007-60,SP2007-155,WIT2007-60, pp. 31–36, 2008.

[4] 根本雄介, 河原達也, 秋田祐哉. スライド情報を用いた言語モデル適応による講義の音声認識と字幕付与. 2007-SLP-66-16, 2007.

[5] 安藤彰男, 今井亨, 小林彰夫, 本間真一, 後藤淳, 清山信正, 三島剛, 小早川健, 佐藤庄衛, 尾上和穂, 世木寛之, 今井篤, 松井淳, 中村章, 田中英輝, 都木徹, 宮坂栄一, 磯野春雄. 音声認識を利用した放送用ニュース字幕制作システム. 電子情報通信学会論文誌, Vol. J84-DII, No. 6, pp. 877–887, 2001.

[6] 池谷友秀, 林貴宏, 尾内理紀夫. 吹き出し形式の字幕表示システム:なかじまくん. 電子情報通信学会技術研究報告, MVE2007-36,ISSN:09135685, 2007.

[7] 宮崎観世, 瀬川典久, 阿部芳彦, 村山優子. ビデオ会議における発言表示手法の提案. 電子情報通信学会技術研究報告, ISO2004-22, 2004.

[8] 前野昌弘, 三國彰. 図解でわかる統計解析. ISBN-534-0303603. 日本実業出版社, 2000.

[9] 李晃伸. 大語彙連続音声認識エンジン Julius ver.4. 情報処理学会研究報告, 2007-SLP-69-53, 2007.

[10] 河原達也, 李晃伸. 連続音声認識ソフトウエア Julius. 人工知能学会誌, Vol. 20, No. 1, pp. 41–49, 2005.

[11] 奈良先端科学技術大学院大学 OpenCV プログラミングブック制作チーム. OpenCV プログラミングブック. ISBN-978-4-8399-2354-9. 株式会社毎日コミュニケーションズ, 2008.