画像の直線検出に基づく音声中の検索語検出のための画像用フィルタ

則武和幸[†1]　南條浩輝[†1]　吉見毅彦[†1]

講演などの音声から，聞きたい検索語が出現する区間を特定する検索語検出の研究を行う．音声中の検索語検出では，検索対象の音声を音声認識システムでテキスト化する必要があり，音声認識誤りへの対応が重要となる．これまでに，画像中の直線検出に基づく検索語検出がある．これは，検索語と検索対象の音声認識結果に含まれるサブワード（音節など）の類似度を検索語-音声認識結果画像の各画素にマッピングした画像には，検索語が出現する区間に直線が表れる性質を利用したものである．しかし，この手法では音声認識における削除誤りや挿入誤りに対応できないという問題がある．本研究ではこれらの誤りに対応するための直線強調用および画像中の雑音除去用の画像用フィルタを研究する．未知語の検出タスクにおいて直線強調用フィルタにより再現率の大きな向上が得られ，雑音除去用フィルタと検索語長ごとのしきい値の変更により適合率の向上も得られた．

Line Detection-oriented Image Processing Filters for Spoken Term Detection

Kazuyuki NORITAKE,[†1] Hiroaki NANJO[†1]
and Takehiko YOSHIMI [†1]

Spoken term detection (STD) from oral presentations is addressed. Specifically, we regard a STD as a line detection problem from an image file, in which each pixel holds a syllable-distance between query term and automatic speech recognition (ASR) results. Since such kind of image file essentially includes ASR errors, line detection from noisy image should be investigated. In this paper, we propose line detection-oriented image processing filters for STD. We achieved 0.46 of F-measure for low frequency term (out of vocabulary term in ASR system) detection task, and 0.75 of F-measure for known term (in-vocabulary term in ASR system) detection task.

1. はじめに

近年，デジタル化されて保存されている音声や動画が増加している．これに伴いこれらの大量のデータから見たい，聞きたい部分を検索したいという機能が求められるようになった．音声を含むデータに対しては，音声認識技術を適用してデータを検索するという方式が有望である．特に音声中に検索語が出現する区間を特定する問題は音声中の検索語検出（Spoken Term Detection: STD）と呼ばれ，盛んに研究が行われている[1)2)3)4)]．

STD の手法の一つに画像の直線検出手法であるハフ変換を適用した研究[5)] がある．具体的には，検索語を構成する各サブワード（音素や音節など）を縦軸に，音声認識結果を構成する各サブワードを横軸にとって，各画素に検索語の各サブワードと音声認識結果の各サブワード間の距離を濃度として与える．このように作成した画像には検索語が含まれいる区間に直線が表れるため，その直線をハフ変換に基づいて検出することで STD が行える．ワードスポッティングにもこれに似た手法がある[6)]．しかし，これらの手法には問題がある．具体的には，音声認識において，削除誤りや挿入誤りが発生した場合には，画像中に直線が正しく表れないため，検索語を検出できないことがある．そこで，本研究ではこの削除誤りと挿入誤りに対応するため，直線検出を行う前に画像に対して画像処理（フィルタリング）を行い，削除誤りや挿入誤りが発生しても直線検出を検出しやすい画像を作成する．これにより音声認識誤りに対して柔軟な対応ができる検索語検出法を実現する．

本論文の構成について述べる．2 章では音声中の検索語検出の一般的な説明を行い，先行研究であるハフ変換に基づく手法およびその問題点を述べる．3 章では本研究で提案する画像に対して用いるフィルタについて述べる．4 章では評価実験について述べる．5 章では結論を述べる．

2. 画像の直線検出法に基づく検索語検出

はじめに，画像の直線検出法に基づく検索語検出について説明する．図 1 に示すように，検索語の各音節を縦軸に，音声認識結果の各音節列を横軸にとり 2 次元画像を作成する．検索語と音声認識結果の各音節の格子点にはそれらの音節どうしの誤りやすさを反映する何らかの距離尺度，すなわち音節間距離をとる．この音節間距離を画像の濃度値（0 が黒，255

†1 龍谷大学理工学部
Faculty of Science and Technology, Ryukoku University

2

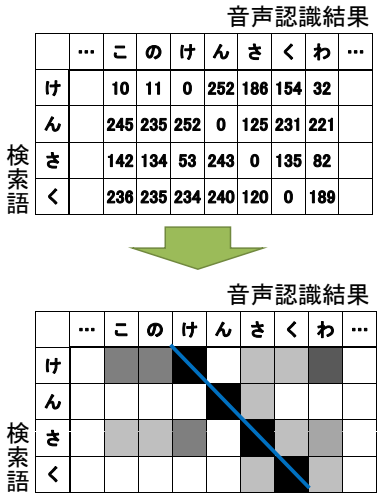

図 1　音節間距離画像に表れる直線

が白）にマッピングさせると，図 1 下部のように音声認識結果の音節列の中に検索語が出現している区間には黒い直線が表れる．このことから STD の問題を画像中の直線検出問題に置き換えることができる．

画像中の直線検出はハフ変換に基づいて行うことができる[7]．STD では直線は斜め 45 度で出現すると仮定できるため，以下の定式化が行える．すなわち，画像の縦の画素数を p，横の画素数を q とし，格子点の画素にあたる音節間距離を $D_{i,j}$ $(1 \leq i \leq p, 1 \leq j \leq q-p)$ として，以下の累積距離 T_j の小さい箇所を求める問題として定式化できる．

$$T_j = \sum_{i=1}^{p} D_{i,i+j-1} \tag{1}$$

検索語検出では，音声認識誤りがある場合には累積距離 T_j が検索語の音節数 p に比例して大きくなる傾向があるので，その絶対値には意味がない．実際には，1 音節あたりの平均累積距離 M_j と適当なしきい値 α（式（2））を用いて直線検出を行う．

$$M_j = T_j / p$$

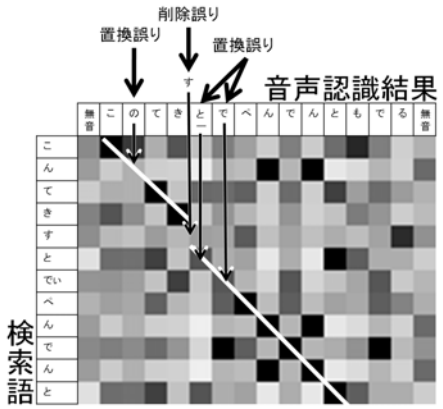

図 2　認識誤りをしている正解区間の例

$$\begin{cases} \text{検索語がある} & \text{if } M_j < \alpha \\ \text{検索語がない} & otherwise \end{cases} \tag{2}$$

次に，この手法の問題について述べる．音声認識の時点で誤りがあると，直線を検出できないことがある．音声認識誤りが含まれる場合の検索対象音声認識結果-検索語画像の例を図 2 に示す．少量の置換誤りに関しては，直線上の一部の画素の濃度値が大きくなるだけであり，ハフ変換の性質上，致命的な問題とはならない．これに対して挿入誤り，削除誤りが存在する場合は問題が大きい．図 2 に示すとおり，削除誤りが存在すると直線が下に直線がずれる．挿入誤りが存在する場合は直線が右にずれる．この結果，直線検出の平均累積距離の計算（式（2））では本来検索語がある位置に直線を検出できない．

3. 音声認識誤りに対応するための画像用フィルタ

3.1 直線強調フィルタ

ここでは，2 章で述べた画像の直線検出法に基づく検索語検出の問題を解決するために用

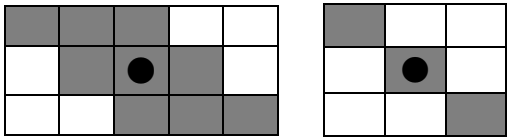

図 3　左：直線強調フィルタ，右：雑音除去フィルタ

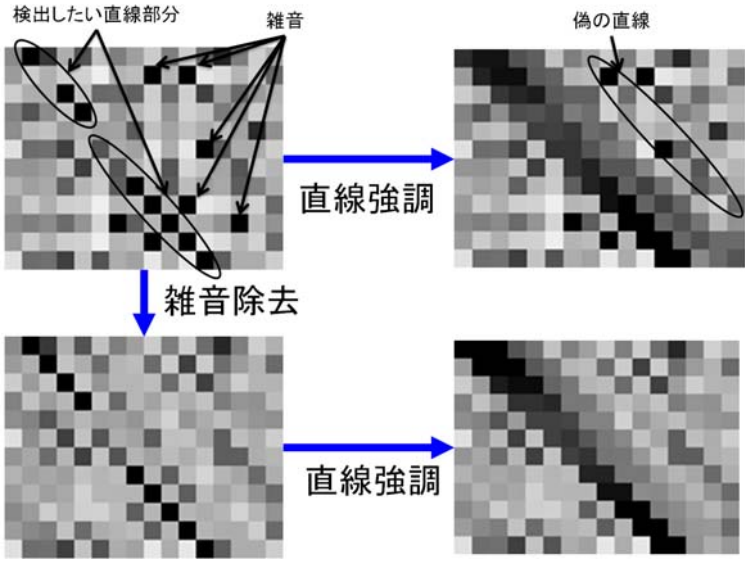

図 4　フィルタリングによる画像の変化

いる直線強調フィルタについて述べる．本研究では削除誤り，挿入誤りに対応することを目的にフィルタを提案する．具体的には，図 3 の左側に示すフィルタを提案する．これは，中心（図 3 の黒丸）の画素値を灰色の範囲内の画素のうち最も黒い（濃度値の近い）3 つの画素の画素値の平均値に置き換えるフィルタである．このフィルタは『灰色の範囲内に直線があるとき，削除誤りや挿入誤りが発生していても灰色の範囲内の 3 点には直線に含まれる画素が含まれる』という仮定に基づいている．実際に，このフィルタをかけた例を図 4 の上段に示す．直線がある区間に太い直線が表れており，削除誤りや挿入誤りがある区間でも直線を検出することができるようになったことがわかる．しかし，このフィルタをかけると画像中の雑音（周辺画素の値と著しくかけ離れた画素）が多い区間でも直線を検出してしまうことがある（例：図 4 の上段右側）．この問題を解決するため次の 3.2 節で述べる雑音除去フィルタを用いる．

3.2　雑音除去フィルタ

次に，画像的な雑音である点画像を除去するための雑音除去フィルタについて述べる．画像処理における雑音除去では特徴抽出を行う前に平滑化フィルタやメディアンフィルタ[7] にかけることが一般的である．本研究で扱うような画像においては，平滑化フィルタやメディアンフィルタをかけると本来の直線も除去してしまう．このことから，斜め 45 度方向の直線を保持したまま雑音を除去できるようメディアンフィルタの参照範囲を少なくしたフィルタを検討する．このフィルタは図 3 の右側に示されている．これは，中心の画素の値を灰色部分の画素値の中間値に置き換えるものである．このフィルタをかけると図 4 の下部左側に示す通り黒い孤立点が薄くなるため，その後の直線強調フィルタ処理においても雑音除去しないときに比べて，直線が存在する付近以外の画素の濃度が薄くなっていることがわかる（図 4 下部右側）．

4.　評 価 実 験

4.1　実 験 条 件

本研究の評価には音声ドキュメント処理ワーキンググループによって作成された Spoken Term Detection のためのテストコレクション（以下，テストコレクション）[8][9] を用いる．検索対象は，コア講演セット 177 講演，約 44 時間を音声認識したものを用いる．検索語として，コア講演用既知語セット 50 検索語と，コア講演用未知語セット 50 検索語を用いる．検索単位は各発話区間（200msec で区切られた音声区間）であり，1 つ以上の検索語を含む発話区間が抽出できれば正解とする．

4.2　検索システム

はじめに，本研究で用いた検索システムについて述べる．検索システムのアルゴリズムを図 5 に示す．長さ p の検索語が入力されると，検索語と k 番目 $(k = 1 \ldots N)$ の発話の音声認識結果から距離画像を作成する．次に，その中に検索語があるかをチェックする．具体的には k 番目の発話 U_k の長さが q のとき，式（2）に基づき $M_j < \alpha$ となる $j(1 \leq j \leq q-p)$ があれば U_k に検索語があるとする．検索語がなければ画像に対してフィルタ処理を行い，再度直線検出を行う．検索語があれば，U_k に検索語があるとする．これを発話が終わるまでくり返し行い $(k = 1 \ldots N)$，検索語を含む発話のリストをユーザーに返す．

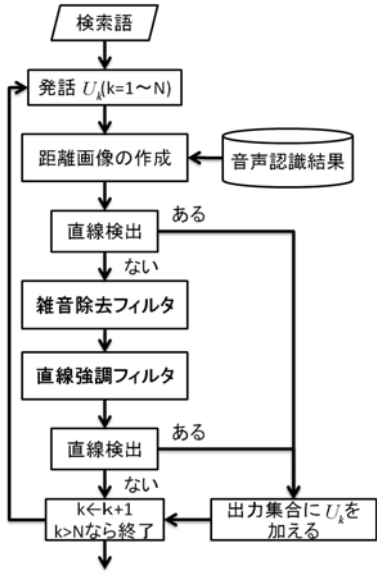

図 5　検索システムのアルゴリズム

次に，画像を作成する際に用いる距離尺度について述べる．距離尺度は何でもよいが，今回は各音素 HMM（Hidden Markov Model）のバタチャリヤ距離[10] を用いた．今回使用した HMM は書籍 11) 付属の 1 混合 monophone 男性用の音響モデルである．この音響モデルは日本音響学会の音素バランス文からなる研究用連続音声データベース（ASJ-PB）の全部と，新聞記事読み上げ音声コーパス（ASJ-JNAS）のうち 100 名分を学習に利用したものである．この音素 HMM は初期状態と終了状態を含め 3 つの状態から構成されている．各状態は 25 次元ガウス分布を保持している．第 1 状態，第 3 状態は前後の音節に依存した値となるので，第 2 状態のみを用い，これをその音素の分布とみなす．第 2 状態同士のガウス分布のバタチャリヤ距離 BD は式（3）で定義される．ここで $\mu_{\mathbf{1}}$ と $\mu_{\mathbf{2}}$ は各音素の特徴ベクトルの平均，Σ_1 と Σ_2 は分散共分散行列である．

$$BD = 1/8(\mu_{\mathbf{2}} - \mu_{\mathbf{1}})^t \left[\frac{\Sigma_1 + \Sigma_2}{2}\right](\mu_{\mathbf{2}} - \mu_{\mathbf{1}}) + \frac{1}{2}\ln\frac{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}} \tag{3}$$

画像の濃度値として使うためにこの BD を 0 から 255 の範囲に正規化する必要がある．したがって本研究では音素間距離 D_p を式（4）と定義する．

$$D_p = 255(1 - e^{\beta BD}) \qquad (0 \le D_p \le 255) \tag{4}$$

ここで β は調整のための係数である．β を大きくすると BD が 0 以外の値のとき D_p は 255 に近い値をとり，β を小さくすると BD が 0 以外の値のとき D_p は 0 に近い値となる．本実験では，D_p がなるべく均等に分布するようにこの β を実験的に 0.75 に決定した．次に，この音素間距離を用いた音節間距離について述べる．日本語の音節は子音と母音に分割できるため，音節 C_1V_1 と C_2V_2 の音節間距離 $D_s(C_1V_1, C_2V_2)$ は，それぞれの子音の音素間距離 $D_p(C_1, C_2)$，母音の音素間距離 $D_p(V_1, V_2)$ の平均として定義する（式（5））．

$$D_s(C_1V_1, C_2V_2) = \frac{D_p(C_1, C_2) + D_p(V_1, V_2)}{2} \qquad (0 \le D_s \le 255) \tag{5}$$

なお，母音の「あ」や撥音の「ん」など 1 音素 P のみから構成される音節に関しては，$C = P$，$V = P$ として式（5）に基づき計算する．

4.3　画像用フィルタの評価結果

はじめに，画像用フィルタの評価を行った．フィルタを用いない検索，直線強調フィルタのみを用いた検索，直線強調フィルタと雑音除去フィルタを組み合わせて用いた検索の評価を行った．再現率と適合率は 1 検索語ごとに算出し，これらを全 50 検索語で平均をとった．しきい値 α を 0 から 80 まで 10 きざみで変化させたときの 50 検索語の平均再現率，適合率，F 値をそれぞれ，図 6 に示す．

フィルタを用いない場合のしきい値 0 の結果はテキストベースで完全一致をとりだすのと同じ処理である．このときの再現率は 0.05 であり，ほとんど検出できていないことがわかる．この結果は，音声認識結果に対する単純なテキスト検索では未知語を高い精度で検索できないことを示している．

直線強調フィルタのみを用いた場合は，フィルタを用いない場合に比べて再現率は向上しており，音声認識誤りがある場合でもより多くの正しい検索語位置をみつけられるようになっていることがわかる．ただし，適合率が著しく低下しており，もともと検索語がないと

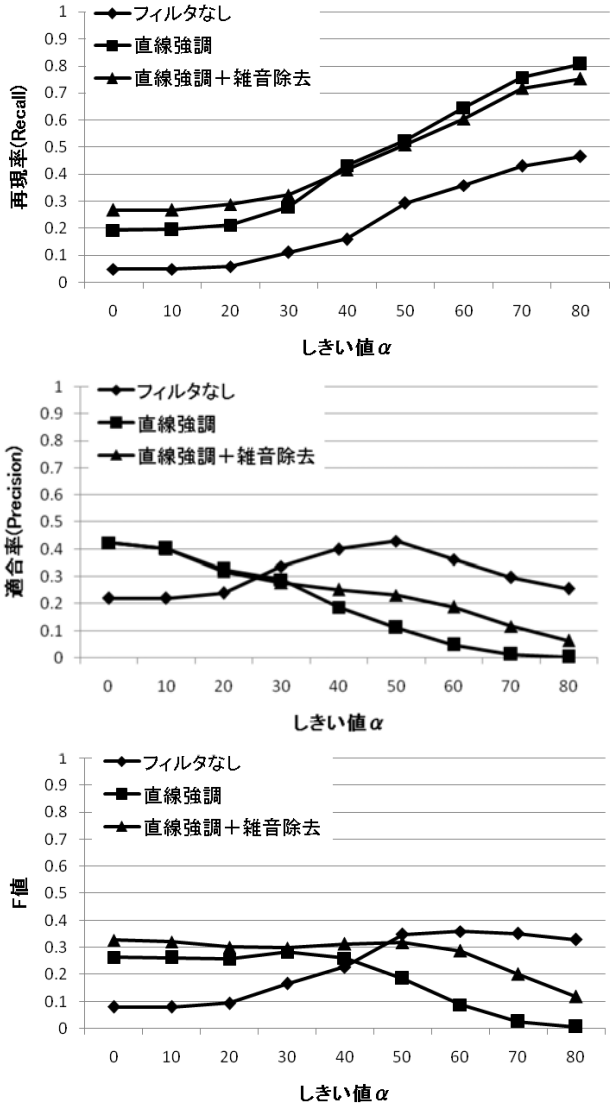

図 6　未知語検出の評価

表 1　未知語の評価結果の比較

	再現率	適合率	F 値
完全一致	0.05	0.22	0.08
連続 DP	0.38	0.22	0.28
フィルタなし（しきい値を 50 に固定）	0.36	0.36	0.36
提案手法（しきい値を 50 に固定）	0.51	0.23	0.32
提案手法（しきい値を検索語長により変更）	0.49	0.42	0.46

表 2　提案手法の検索語長ごとの結果

検索語長	再現率	適合率	しきい値
6 音節未満	0.25	0.57	0
6 音節以上 8 音節未満	0.44	0.34	60
8 音節以上	0.86	0.4	70
平均	0.49	0.42	

ころも誤って検出していることがわかる．直線強調フィルタと雑音除去フィルタを用いた場合は，直線強調フィルタのみを用いた場合と比べると適合率を改善できている．例えば，しきい値 50 のときで比較すると，再現率は 0.51 から 0.52 と，ほぼ同等の値を保ちつつ，適合率を 0.11 から 0.23 と改善できていることがわかる．この結果は雑音除去により，不要なわき出し誤りを抑えつつ正しい箇所を検出できることを示している．

フィルタなしの結果（しきい値を 50 に固定），提案手法の結果（しきい値を 50 に固定），連続 DP マッチングで音節認識誤りを 2 つまで許容した結果を表 1 に示す．連続 DP マッチングと画像中の直線検出（フィルタなしの結果）を比べると，再現率は同等であるが適合率が高い．この差は距離尺度の違いによるものと考えられる．連続 DP では音節間の距離を 0-1 の 2 値としているのに対して，画像中の直線検出では 0 から 255 の音節間距離を用いており，その効果と考えられる．

次に，誤検出について詳細に調べたところ，検索語長の短い検索語において誤検出が比較的多いことがわかった．そこで，検索語を検索語長で 3 つに分割し，それぞれ式（2）のしきい値 α として異なるものを用いて検索した．具体的には，6 音節未満，6 音節以上 8 音節未満，8 音節以上の 3 つのグループにわけ，それぞれの F 値が最適となるしきい値を用いた．実際に，6 音節未満の検索語に対してはしきい値が 0，6 音節以上 8 音節未満の検索語に対してはしきい値が 60，8 音節以上の検索語に対してはしきい値が 70 のときに F 値が最も高くなった．結果を表 2 に示す．また他手法との比較のために表 1 の下段にも示されて

表 3　既知語の評価結果の比較

	再現率	適合率	F 値
完全一致	0.56	0.95	0.7
連続 DP	0.56	0.95	0.7
フィルタなし（しきい値固定）	0.66	0.83	0.74
提案手法（しきい値固定）	0.62	0.87	0.72
提案手法（しきい値を検索語長により変更）	0.66	0.87	0.75

いる．

長い検索語では高い再現率を得られていることがわかる．これに対して短い検索語に関しては再現率が低いことがわかった．しきい値を大きくすることで再現率は向上できるものの適合率の低下が大きく F 値が下がることも確認できた．短い検索語に対して誤検出を抑制する必要があることがわかった．

長さによりしきい値を変更したときと，しきい値を固定したときは，再現率はともに約 0.5 であった．これは，音声中の聞きたい区間を約半分見つけることができることを示している．適合率についてはしきい値を可変とすることで 0.23 から 0.42 に大きく向上した．これは，検出した区間のリストのうち 10 件中 4 件は正しいデータであることを示しており，ユーザーが実際に音声を聞いて確かめる STD アプリケーションを考えた場合に，システムを使わない場合に比べて十分はやく欲しい情報にアクセスできる精度であるといえる．

参考のために既知語に関しても未知語と同じ方法で実験を行った．完全一致するもののみを検出した結果（単純なテキストマッチング），フィルタなしの結果（しきい値 30 に固定），提案手法の結果（しきい値を 10 に固定），提案手法の結果（しきい値を変更 0，10，40）を表 3 に示す．既知語に関しては完全一致のみを検出するだけでも多くの検索語を検出できることがわかる．連続 DP マッチングでは再現率を向上させるために音節認識誤りを許容すると，適合率が著しく低下する．今回の実験では誤りを許さない連続 DP マッチング（完全一致のみを検出）が最適な F 値となった．未知語対象の STD と比べると改善率は小さいものの，提案法により再現率を下げることなく適合率を 0.83 から 0.87 に改善でき，提案法は既知語に対しても頑健に動作することを確認した．

5. おわりに

画像の直線検出に基づく音声中の検索語検出のための画像用フィルタを設計し，その画像用フィルタを用いて画像の直線検出に基づく音声中の検索語検出の評価を行った．未知語セットに関しては，提案手法（フィルタ処理）により再現率の向上が得られたものの，適合率が低下した．検索語の音節の長さで異なる検出しきい値を用いることで，再現率を同等に保ちつつ適合率を 0.23 から 0.42 に大きく向上できた．

本手法は短い検索語の検出に問題があるので，短い検索語に関しては他の手法を組み合わせて用いた手法を考えることを行っていきたい．

参 考 文 献

1) 澤田心太，桂田浩一，手島茂樹，入部百合絵，新田恒雄：大規模音声ドキュメントからの高速キーワード検索法の提案とその評価，日本音響学会講演論文集, 秋季，2-9-10 (2010).
2) 松永　徹，趙　國，山下洋一：音声ドキュメント検索語検出における音響情報を用いた再評価，日本音響学会講演論文集, 秋季，2-9-12 (2010).
3) 名取　賢，西崎博光，関口芳廣：音声中の検索語検出のための複数の音声認識結果を用いたネットワーク型インデキシング，日本音響学会講演論文集, 秋季，2-9-9 (2010).
4) 伊藤慶明，西崎博光，胡　新輝，南條浩輝，秋葉友良，相川清明，河原達也，中川聖一，松井知子，山下洋一：音声中の検索語検出のためのテストコレクション構築–中間報告–，情報処理学会研究報告，SLP-78-4 (2009).
5) 金子泰輔，秋葉友良：ハフ変換に基づく音声ドキュメントの高速検索語検出法，日本音響学会講演論文集, 春季，3-6-10 (2010).
6) 西　宏之，木村義政，グエン・ヴァン・ドン：距離マトリックス画像のハフ変換を用いたワードスポッティング，日本音響学会講演論文集, 春季，1-5-1 (2009).
7) 内村圭一，岩崎洋一郎，松島宏典：画像処理入門，培風館 (2010). ISBN:978-4-563-01583-1，pp82-91.
8) 西崎博光，胡　新輝，南條浩輝，伊藤慶明，秋葉友良，河原友良，中川聖一，松井知子，山下洋一，相川清明：Spoken Term Detection のためのテストコレクション構築とベースライン評価，情報処理学会研究報告，Vol.2010-SLP-81 No.13 (2010).
9) Itoh, Y., Nishizaki, H., Hu, X., Nanjo, H., Akiba, T., Kawahara, T., Nakagawa, S., Matsui, T., Yamashita, Y. and Aikawa, K.: Constructing Japanese Test Collections for Spoken Term Detection, *11th International Congress on Spoken Language Processing (INTERSPEECH 2010 ICSLP)*, pp.677-680 (2010).
10) O.Duda, R., E.Hart, P. and G.Stork, D.: *Pattern Classification*, Wiley-interscience (2000). ISBN:0-471-05669-3，pp45-50.
11) 鹿野清宏，伊藤克亘，河原達也，武田一哉，山本幹雄：音声認識システム，オーム社出版局 (2001). ISBN:4-274-13228-5，pp17-23，131-146.