社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2006－SLP－64 (37)
2006／12／22

講義音声認識における収録装置とケプストラム正規化法の検討

富樫慎吾† 北岡教英†† 中川聖一†

† 豊橋技術科学大学 情報工学系
†† 名古屋大学大学院 情報科学研究科
E-mail: †{togashi,kitaoka,nakagawa}@slp.ics.tut.ac.jp

あらまし 本研究では，講義情報をセマンティックコンテンツ化するための要素技術である音声認識による自動書き起こしの精度向上に関する検討を行った．まず，複数のマイク・音響モデル・デコーダを種々組み合わせて認識実験を行い，環境によって認識精度にどの程度影響があるのかを調べた．この結果，マイクの違いが認識精度に与える影響が非常に大きく，ハンドマイクで収録した音声と，ワイヤレスピンマイクを用いて圧縮処理を行った音声では，認識精度に最大で約23%の差があらわれることがわかった．しかし，収録時の利便性を考えると，実用上はワイヤレスピンマイクの方が望ましい．そこで，こうした粗雑な収録環境で得られた音声からも高精度に音声認識を行うために，ケプストラムの正規化手法として平均正規化 (CMN)、分散正規化 (CVN)，そしてケプストラムの分布を一致させるヒストグラム正規化 (HEQ) を実験し，またその実行単位についても検討を行った．
キーワード 講義音声、音声認識、ケプストラム正規化，ヒストグラム正規化

An Investigation of Variation of Recording Systems and Cepstrum Normalization for Recognition of Lecture Speech

Shingo TOGASHI†, Norihide KITAOKA††, and Seiichi NAKAGAWA†

† Toyohashi University of Technology
†† Nagoya University
E-mail: †{togashi,kitaoka,nakagawa}@slp.ics.tut.ac.jp

Abstract In this paper, we described the investigation of methods to improve the performance of ASR used for extending lecture information into semantic contents. First, we investigate the effects of difference among microphones, acoustic models and decorders. It turned out that using pin-microphone decrease up to 23% compared to handheld-microphone in accuracy. There is significant impact by the difference between handheld-microphone and pin-microphone to performance. To improve the performance even when using a cheap equipment like this, we tried to apply some cepstrum normalization methods, like CMN, CVN and HEQ to lecture speech data, in which we considered about unit of the normalization process.
Key words Lecture Speech, Automatic Speech Recognition, Cepstrum Normalization, Histogram Normalization

1. まえがき

近年，インターネット上の映像・音声コンテンツは年々増加しつつあり，ユーザがその膨大なデータ群の中から必要とする情報を検索する作業はしだいに煩わしいものとなってきている．そこで，それらのデータに対して，発話内容を計算機が自動的に理解し，処理できるようにアノテーションを行うセマンティックコンテンツ化の必要性が高まっている．また，こうした技術を応用して，講義や講演をビデオ録画し、自宅からでも容易に学習や復習が可能なシステムが実用化されている [1]。我々の研究室でも，収録した講義音声／動画をネットワークを通じて配信し，同時にスライド情報や自動書き起こし情報、およびそれを元にした自動要約情報や，キーワードインデックス情報なども利用可能にする，講義音声の高度利用学習システムを開発中

である [2] [3].

本研究では，講義情報をセマンティックコンテンツ化するための要素技術である音声認識による自動書き起こしの精度向上に関する検討を行った．講義音声においては，あいまいで不明瞭な発音も含んだ話し言葉が主となる上，周囲の雑音の影響も考慮する必要があるなど，高精度の認識を阻むいくつかの問題が存在する．本研究ではまず，複数のマイク・音響モデル・デコーダを種々組み合わせて認識実験を行い，環境によって認識精度にどの程度影響があるのかを調べた．この結果，マイクの違いが認識精度に与える影響が非常に大きく，指向性ハンドマイクで収録した音声と，ワイヤレスピンマイクを用いて圧縮処理を行った音声では，認識精度に最大で約 23%の差があらわれることがわかった．しかし，実際に行われている講義や会議の録音では，多くはピンマイクなどの簡素な機材を使用しており，取扱いの簡便さや話者の負担などの観点からみても，高性能な指向性ハンドマイクや無圧縮の記録方式を使用することは実用的ではない．そこで，こうした粗雑な収録環境で得られた音声からも高精度に音声認識を行うために，ケプストラムの正規化手法として平均正規化 (CMN)、分散正規化 (CVN)，ヒストグラム正規化 (HEQ) を実験し，またその実行単位についても検討を行った．最後に，我々が使用しているコンテキスト独立の音節単位の音響モデルを，コンテキスト依存モデルにした場合についても検討した．

2. 講義音声収録

収録対象とする講義は、我々の大学で実施されている大学院／学部向けのものである。講義は一回 150 分で，前後半 75 分ずつとなっている。現在までに 6 名の講義を合計 8 回分収録した。その際、録音機材が認識結果に与える影響を調査するため、各講義について、表 1 に示す 3 種類の装置を同時に使用して収録を行った。

認識対象とする音声は，3 話者／4 講義である．これらの音声データを便宜上表 2 に示すように記述する。1-1 と 1-2 の講義内容は、「音声言語処理の概要」、2-1 と 2-2 は「DP マッチングと連続音声認識」、3-1 と 3-2 は、「音声対話」、4-1 と 4-2 は、「音声認識とパターン認識」である。各講義について、録音機材の違いによって表 1 のソース A, B, C が存在する。以降は、たとえば 1-1 のソース A なら、1-1A と表記する。後に示す結果のグラフでは，1-1A と 1-2A の平均を単に 1A と表記する．なお、表 2 の "PP" は後述する CSJ 言語モデル (語彙サイズ 17K) に対するテストセットパープレキシティを示す。文は、便宜上おおよそ 200ms の無音区間のところで切り出した単位としている。

3. 講義音声認識

2. 節に示した各講義音声について、音声認識実験を行った。認識条件などを以下に記す。

3.1 認識条件

認識率の評価には最初の 10 分程度だけを使用した。これは使用した単語辞書にあわせて、単語分割を人手作業により行う必要があったためである。

この音声試料を、Julius と SPOJUS [5] の 2 種類のデコーダを用いて音声認識実験を行い，結果を比較した．音響モデル及び言語モデルの学習用データは CSJ(日本語話し言葉コーパス 2004 年度版) に収録されている音響学会講演と模擬講演であり，音声認識および対話といったトピックが主であるため，今回の認識対象である講義の内容とドメインは近く、表 2 に示すように未知語率 (語彙サイズ 17K) は比較的小さい。音声認識システムは以下に示すものを用意した。

(a)SPOJUS

SPOJUS は，16kHz でサンプリングされた音声より導出された MFCC(12)、Δ MFCC(12)、Δ POWER(1) の計 25 次元

表 1　データ収録条件

収録データ	マイクロフォン	録音装置
ソース A	SONY C-355 ハンドマイク 周波数特性：20Hz～20,000Hz 指向特性：単一指向性	DAT
ソース B	SONY ECM-C10 ピンマイク 周波数特性：50Hz～15,000Hz 指向特性：全指向性	DAT
ソース C	TOA WM-1300 ワイヤレスピンマイク 指向特性：全指向性	PC (WMV 圧縮)

表 2　認識対象講義音声の概要

表記	話者	総時間	総文数	PP	未知語率
1-1	SN	1:07:56	742	186.4	0.37%
1-2		0:56:59	709	443.7	2.15%
2-1		1:06:11	831	159.8	0.70%
2-2		1:15:28	798	305.5	1.65%
3-1	NK	1:05:49	680	177.7	1.88%
3-2		1:11:14	1099	180.8	3.14%
4-1	TN	1:10:16	582	285.6	1.94%
4-2		1:18:30	648	239.5	2.11%

表 3　認識システム

表記	デコーダ	音響モデル	言語モデル
システム i	SPOJUS	音節 (133 音節)*1)	CSJ*2)
システム ii	Julius	音節 (133 音節)*1)	CSJ*2)
システム iii	Julius	音素 (triphone*3))	CSJ*2)

*1 CSJ 最終版 797 講演 (男性話者) より学習、コンテキスト独立

2 CSJ 最終版で学習、trigram 言語モデル (17635 語彙)

3 CSJ 最終版 DVD に収録。男性 787+女性 166 講演で学習

の特徴ベクトルを使用する。2 パスデコーダであり、1 パス目は HTK ツールキットを用いて学習したコンテキスト独立の 133 音節からなる 5 状態（無音は 3 状態），32 混合（対角共分散行列）の音響モデル [4] と bigram 言語モデルを用い、得られた N-best 候補を trigram でリスコアする [5]。実験では，N=200 とした．

(b)Julius3.5

Julius では特徴パラメータは、上記したものと同じ MFCC, Δ MFCC, Δ POW の 25 次元を用いる．音響モデルは上記と同一のコンテキスト独立音節 HMM，および 3 状態，16 混合 (対角共分散行列) の triphone である．なお，Julius には高速版と高精度版が存在する．高精度版はビームサーチのアルゴリズムが異なり，2-pass 目の単語間 triphone をより厳密に計算するようになる．本実験では高精度版を使用した．

3.2 認識結果

認識精度 (accuracy) の評価結果を図 1 に、正解率 (correct) の評価結果を図 2 に示す(注1)。ここに示すスコアは各講義の前半と後半の平均である (例：1-1A と 1-2A の平均=1A)。なお、集計時にフィラー (全発話単語の 5%前後) や言い淀みなどを取り除く処理は特に行っておらず、それらも認識精度を左右する要因となっている。

マイクによる差という観点からソース A,B,C を比較すると、各講義において，A > B > C という順に認識性能が劣化していることがわかる．人間の耳では A と B の音声はいずれも聞き取りやすい音質でほとんど差はなかったが、特に話者 SN の認識精度には最大 18.8%もの差が表れている。ソース C は，マイクの性能としてはソース B のものと同程度であるが，録音時に WMV による圧縮をかけるために，かなり機械的な音声に変質しており，ノイズも増幅されているような印象を受けた．結果として，ソース B と C の間には最大 15.8%の認識精度の差が表れている (話者 NK)．さらにソース A と C では，話者 TN で 23%もの認識精度の差が表れた。

デコーダの比較という観点からシステム i と ii を比較すると，さほど差は顕著ではないが，正解率 (correct) はシステム ii の方がやや高く，正解精度 (accuracy) は講義によってバラツキがあるものの，全体としてシステム i の方が良いということが言える．また，話者による認識性能の差という観点では、NK の講義音声が全体的に認識性能が高い．これは、話者 NK が音響モデル作成話者集合に近い年齢であったためと発声が比較的明瞭であったためと思われ、使用するシステムによって認識が容易／困難な声の特性というものが存在することを示している。

音響モデルの比較という観点からシステム ii と iii を比較すると，学習データが若干異なるので厳密な比較はできないが、正解率では同等，認識精度ではトライフォンモデルが音節モデルを上回っている場合の方が多かった。音節モデルのコンテキスト依存化が必要と思われ [6]，次節の実験を行った．

3.3 コンテキスト依存音節モデルによる認識

本研究室で作成したコンテキスト依存 (CD) 音節モデルは，4.1 に記述するコンテキスト独立 (CI)116 音節モデルを種モデルとしている．各音節に対して，"N"，"⟨SIL⟩"(文頭，文末の無音)，"⟨sp⟩"(文中の短い無音)，および 5 母音が直前に接続する際の変音を考慮して学習し，928 個の HMM を作成した．なお，直後の分節音に対する依存は考慮していない．学習データは CSJ 最終版の，男性話者 814 講演である．状態数，混合数は 116 音節モデルと同様 3 状態，4 混合である．

このモデルを，4.2 節に示す一発話ごとの CMN を行った認識結果 (200 ベスト) に対するリスコアに用いた．時間の制約上，今のところ得られているのは話者 SN の一回目の講義の前半 (1-1) のリスコア結果のみである．結果を図 3，4 に示す．

ソース A に対しては約 1%，ソース B に対しては約 2%の認識率の向上があったが，ソース C に対してはコンテキスト依存の効果がなく，逆に悪化している．この現象は，ワイヤレスピンマイクのような音声波形が歪む劣悪な環境下で起こるものと考えられる．テストデータを増やして確認の実験を行う必要がある．なお，ソース B がソース C より認識結果が悪いのは，8 個の講義音声ソースのうち，1-1 だけであった．

4. ケプストラム正規化法

収録した音声を認識する際に問題となるのは，認識対象音声と学習用音声との間の収録環境の差異によるパラメータのミスマッチである．このミスマッチを解消するための代表的な手法として，CMN(平均正規化) および CVN(分散正規化) [7] [8]，HEQ(ヒストグラム正規化) [9] など，学習／認識音声のケプス

(注1)：文献 [2] と結果が異なっている理由は、リスコア時の言語重みとペナルティを今回適正に設定したためである．

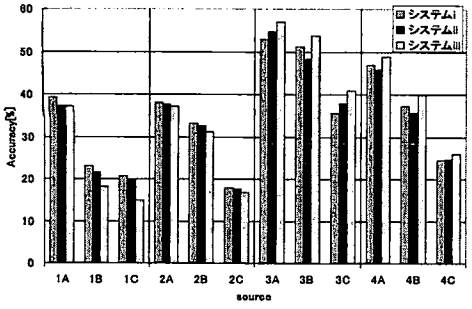

図 1 各種条件による認識結果 (accuracy)

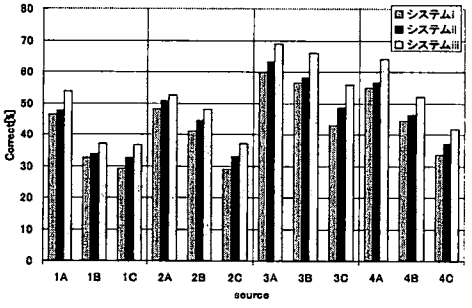

図 2 各種条件による認識結果 (correct)

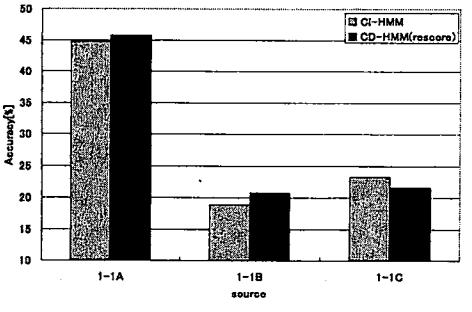

図 3 CDHMM 認識結果 (accuracy)

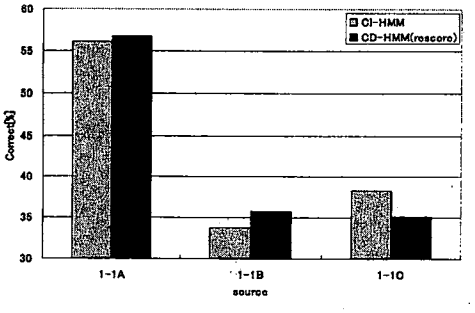

図 4 CDHMM 認識結果 (correct)

トラム値を補正するものがある．本稿ではこれらの手法を表 2 に示した講義音声に適用し，認識精度の向上を図った．

4.1 実験条件

当研究室で使用している音響モデル学習用環境との関係から，学習データおよび音響モデルの仕様は表 3 とは異なっている．

学習用データとして，CSJ 最終版 814 講演 (男性話者) を用い，116 音節のコンテキスト独立音響モデルを作成した．学習プログラムは我々の研究室で作成したもので，状態数は 4，GMM 混合数は 4（全角共分散行列）とした．使用するパラメータは，16kHz でサンプリングされた音声より導出された MFCC(12), ΔMFCC(12), ΔΔMFCC(12), ΔPOWER(1), ΔΔPOWER(1) の計 38 次元である．デコーダには前述の SPOJUS を用いるが，2 パス目の trigram によるリスコアは行っていない．

なお，認識対象の音声は 3.1 節と同様である．

4.2 CMN

CMN による補正は以下の式で表すことができる．

$$\hat{C_t^{(d)}} = C_t^{(d)} - \mu^{(d)} \tag{1}$$

ここで，$C_t^{(d)}$ は時刻 t における d 次のケプストラムである．また，平均 $\mu^{(d)}$ は，CMN の実行単位を一発話単位にするか，それとも複数発話 (一講演／講義) 単位にするかによって変わる．

$$\mu^{(d)} = \begin{cases} \frac{1}{T}\sum_{t=1}^{T} C_t^{(d)} & \text{一発話単位} \\ \frac{1}{T_{all}}\sum_{t=1}^{T_{all}} C_t^{(d)} & \text{講演／講義単位} \end{cases} \tag{2}$$

ここで，T は一発話の全フレーム数を，T_{all} は講演／講義の全発話フレーム数を示す．本稿では一発話単位，および講演／講義単位の両方で，学習用音声と評価用音声共に CMN の処理を行い，結果を比較した．

4.3 CVN

本稿では CVN の実験は CMN とあわせて行った．すなわち，補正は以下の式で表すことができる．

$$\hat{C_t^{(d)}} = \frac{C_t^{(d)} - \mu^{(d)}}{\sigma^{(d)}} \tag{3}$$

ここで，平均 $\mu^{(d)}$ と偏差 $\sigma^{(d)}$ は，式 2 と同様，CMN の実行単位を一発話単位にするか，それとも複数発話 (一講演／講義) 単位にするかによって変わる．

$$\sigma^{(d)} = \begin{cases} \sqrt{\frac{1}{T}\sum_{t=1}^{T} C_t^{(d)2} - \mu_{sent}^{(d)\,2}} & \text{一発話単位} \\ \sqrt{\frac{1}{T_{all}}\sum_{t=1}^{T_{all}} C_t^{(d)2} - \mu_{all}^{(d)\,2}} & \text{講演／講義単位} \end{cases} \tag{4}$$

ここで，$\mu_{sent}^{(d)}$ は一発話のケプストラム値の平均，$\mu_{all}^{(d)}$ は講演／講義のケプストラム値の平均を示す．本稿では一発話単位，および講演／講義単位の両方で CVN の処理を行い，結果を比較した．

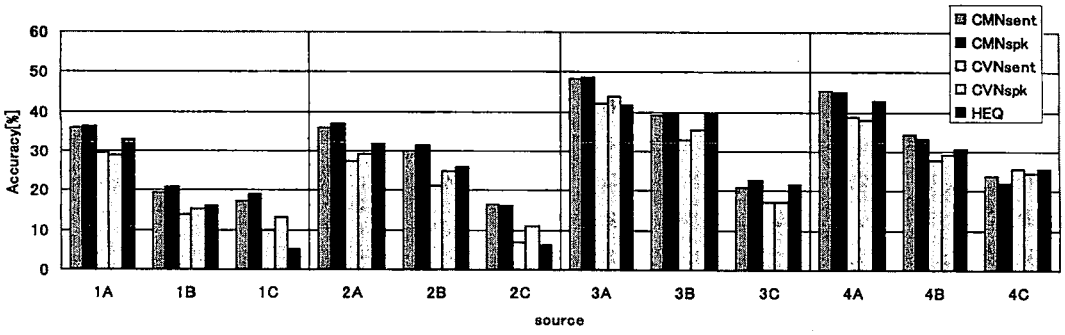

図 5 認識結果 (Accuracy)

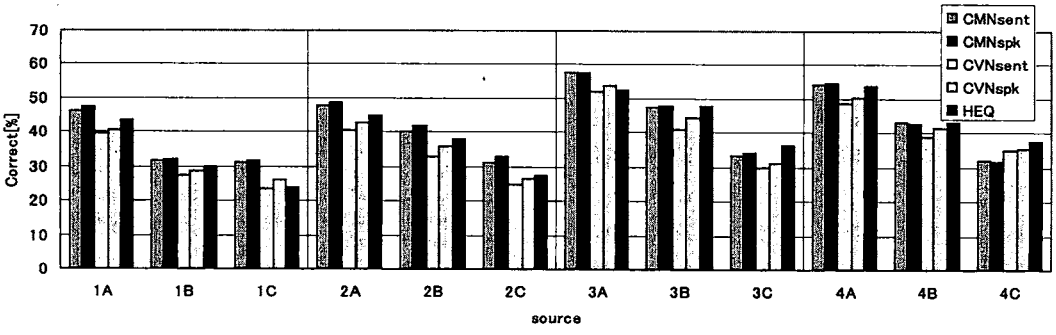

図 6 認識結果 (Correct)

4.4 HEQ

HEQ では，上記の CMN や CVN があくまで分布の平均や分散という一次モーメントしか正規化しない手法であることに対し，確率分布の形状も同一にすることで，異なる話者や収録環境から抽出した特徴量でも見掛け上すべて同一環境から得られたような分布になるため，マイクの差や話者の差に影響されない頑健な認識を行えることが期待できる．HEQ の実行は以下の式で表すことができる．

$$\hat{C_t^{(d)}} = P^{spec-1}(P^{obj}(C_t^{(d)})) \qquad (5)$$

ここで，P^{obj} は適応対象音声の累積分布関数，P^{spec-1} は標本音声の累積分布関数の逆関数である．すなわち，実験手順としては，まず標本音声と適応対象音声の累積分布を特徴ベクトルの各次元ごとに作成する．今回は $\pm 3\sigma$ の範囲でビンを 100 等分して作成した．次に，フレームごとに入力される対象音声のケプストラム値での累積を，対象音声の累積分布から線形補間により求め，標本音声の同一の累積の位置でのケプストラム値をこれも線形補間により求め，この値で対象音声のケプストラム値を置き換える．このようにして対象音声の累積分布を標本音声の分布に一致させることで，ヒストグラムで近似された確率分布を同一の形に正規化することができる．本稿では CSJ 最終版に収録された話者 "A01M0007" を標本音声とし，学習データのヒストグラムを全て正規化してモデルを作成した．また，認識対象音声である各講義のヒストグラムも正規化し，作成したモデルを用いて認識実験を行った．

5. 実験結果

各正規化手法を用いた際の認識率を，図 6, 5 に示す．

ここで，各項目の sent ／ spk は，一発話単位／講義単位であることを表す．

CMN の結果を見ると，実行単位は講義単位の方が，一発話単位よりもわずかながら向上している．長い単位で平均を計算することは，短い発話でうまく平均を推定できないといった状況において有利であることが想像できるが，実際に認識対象の発話を見てみると，極端に短い（1 秒未満）発話は全発話中の一割程度であった．話者は常に動きながら講義をするため，その都度伝達特性が異なってくる可能性がある．そういった場合では一発話単位の正規化の方が有利になると考えられる．

つぎに CVN の結果であるが，実行単位にかかわらず，いずれも平均のみを正規化した場合と比べて結果が悪くなった．現在のところ悪化の原因と思しき要因は特定できていない．講義音声以外のテストデータにも同様の処理を行い，ある特定のタスクでは逆効果となるのかどうかを検証する必要がある．

最後に HEQ の結果であるが，これも講義によっては認識精度が向上しているものの，全体的な性能は CMN にも劣っている結果となった．原因としては，学習音声および認識対象音声中に残っている非発話区間のために，正しく累積分布を推定できていないことが考えられる．現在，音声を分割する際には，パワーの閾値で発話／非発話区間を判断しているので，音声以外のノイズなどの非発話区間が完全には削除されていない可能性が高い．また，今回平均値 $\pm 3\sigma$ の範囲で累積分布を作成し，値を補正したが，その範囲から外れるケプストラム値には補正を加えていないため，$\pm 3\sigma$ の範囲外のケプストラム値が多くなるような分布であれば，HEQ が有効に機能しない可能性もある．実際に HEQ 適用後の分布を調べてみたところ，一部の話者では適応されないままのケプストラム値が多く，形状が若干標本音声と異なってしまっているといった場合が見られた．範囲のとりかたによって結果がどのように変化するかを検討する必要がある．

6. おわりに

本稿では高度利用を狙って講義音声を収録する際に重要となる収録装置による音声認識精度への影響，および CMN,CVN,HEQ といった各種ケプストラムの正規化による認識精度の向上手法について述べた．音声認識においては，指向性ハンドマイクとワイヤレスピンマイク (WMV 圧縮音声) で，話者によっては最大 23%の差が表れることがわかった．マイクの特性の違いや音声圧縮などによって認識性能が悪化することを防ぐため，講義音声に対して各種正規化を行い，CMN,CVN については実行単位を一発話／講義単位の両方で実験して結果を比較した．結果として，CMN は講義単位で行った方が認識精度がやや向上したが，CVN はいずれも CMN に及ばない結果となった．また，HEQ では，講義によってはやや向上が見られたが，全体的にはこれも CMN より悪い結果となった．

今後の課題として，より多くのテストデータでの評価，および HEQ においては累積分布作成範囲が結果に与える影響と，発話／非発話区間を GMM で区別し，非発話区間の影響を受けないように分布を作成することなどを予定している．

謝辞 133 音節の HMM を作成していただいた，信州大学の山本一公助手に感謝します．

文　献

[1] 奥村学、久光徹、増山繁、灘波英嗣、福島孝博、中川祐志、渡部聡彦、江原暉将、和田裕二，"テキスト自動要約 知的活動支援の基本技術として"，情報処理学会誌、Vol.43、No.12、pp1286-1316、2002

[2] 富樫慎吾，山口優，北岡教英，中川聖一，"講義音声の認識・要約・インデックス化の検討"，情報処理学会研究報告，2006-SLP-62-11, pp.57-62, 2006

[3] 藤井康寿，富樫慎吾，山口優，北岡教英，中川聖一．"韻律・表層的言語 情報に基づく重要文抽出による講義音声要約の評価，" 日本音響学会講論 集，2-P-28, pp. 149-150, Sep. 2006.

[4] 池田太郎，山本一公，松本弘，西谷正信，宮澤康永：音節連鎖モデルによる大語彙連続音声認識： 第 5 回音声言語シンポジウム、情報処理学会研究報告、2003-SLP-49-26, pp.151-156, 2003.

[5] 北岡教英，高橋伸寿，中川聖一，N-best 線形辞書検索と 1-best 近似木構造辞書探索の併用による大語彙連続音声認識，電子情報通信学会論文誌，Vol.87-DII，No.3，pp.799-807，2004

[6] 北岡教英、梁頴、中川聖一：Trigram・4-gram と文脈依存音響モデルを用いた 1 パス大語彙連続認識アルゴリズムとその高精度化： 電子情報通信学会、音声技報、SP2006-16、2006.

[7] O.Viikki and K.Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition", Speech Communication, vol.25, pp.133-147, 1998

[8] C.Chen, K.Filali, J.Bilmes, "Frontend post-processing and backend model enhancement on the aurora2.0/3.0 databases", Proc.ICSLP 2002, pp.241-244, 2002

[9] S.Molau, M.Pits, and H.Ney, "Histogram based normalization in acoustic feature space", Proc.of ASRU'01, Trento, Italy, December 2001, pp.21-24