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

2007－ＳＬＰ－69（８）
2007／12／20

マルチストリーム HMM を用いた 特徴量の次元別重み付き話者照合の検討

小島 慎也† 岩野 公司† 古井 貞熙†

† 東京工業大学大学院 情報理工学研究科 計算工学専攻
〒 152–8552 東京都目黒区大岡山 2–12–1
E-mail: †{kojima,iwano,furui}@furui.cs.titech.ac.jp

あらまし 本稿では，音響特徴量の各次元を個別のストリームとしたマルチストリーム HMM を用いた話者照合の耐雑音性について検討する．各ストリームの重みは線形判別分析 (LDA) を基づく手法により推定し，教師なしの条件でストリーム重みを推定した場合や，評価データと異なる雑音条件のデータを重み推定用データとして利用し，評価データの雑音条件に依存しない重みを教師ありで推定した場合について性能の評価を行った．特徴量には，ケプストラム特徴量 (MFCC) とスペクトル特徴量 (SPEC) を用いた．様々な種類の雑音・ SN 比条件の日本語 4 桁連続数字音声による話者照合実験を行った結果，どちらの特徴量を用いても，教師なし条件での推定法によって，SN 比 20，15dB では全ての雑音で誤り率が削減されることが分かった．また，雑音条件に依存しない推定方法によって全ての雑音条件で耐雑音性が向上することが確認された．

キーワード 話者照合，耐雑音，マルチストリーム HMM，線形判別分析 (LDA)，スペクトル特徴量

Speaker verification using multi-stream HMMs with dimensionally weighted feature vectors

Shinya KOJIMA†, Koji IWANO†, and Sadaoki FURUI†

† Department of Computer Science, Tokyo Institute of Technology
2–12–1 Ookayama, Meguro-ku, Tokyo, 152–8552 Japan
E-mail: †{kojima,iwano,furui}@furui.cs.titech.ac.jp

Abstract This paper describes a noise-robust speaker verification method using multi-stream HMMs. In this framework, each dimension of acoustic feature vectors is treated as an individual stream and then each stream is automatically weighted by a Linear Discriminant Analysis (LDA) based technique so as to emphasize noise-robust dimensions. This paper proposes two types of weight estimation using the LDA-based technique: unsupervised estimation and supervised noise-independent-weight estimation. In the latter estimation, the LDA is applied to development data contaminated with various noises different from the noise in testing data. Experiments were conducted using Japanese four-connected-digit utterances in various kinds of noise and SNR conditions. Two kinds of acoustic features, cepstrum domain features (MFCC) and spectrum domain features (SPEC), were used for the experiments. For both features, experimental results show that the unsupervised estimation reduces the equal error rates (EERs) in relatively high-SNR (15-20dB) conditions and the noise-independent-weight estimation improves the noise robustness in all noise conditions.

Key words speaker verification, noise robustness, multi-stream HMM, linear discriminant analysis (LDA), spectral domain feature

1. はじめに

マルチストリーム HMM を用いた話者照合手法では，雑音重畳によって信頼性が低くなった特徴量を有するストリームの重みを相対的に小さくすることで，話者照合の耐雑音性を向上させることが可能である．我々の先行研究 [1] では，ケプスト

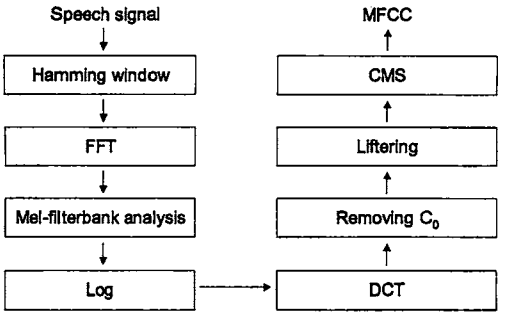

図 1 MFCC の抽出過程

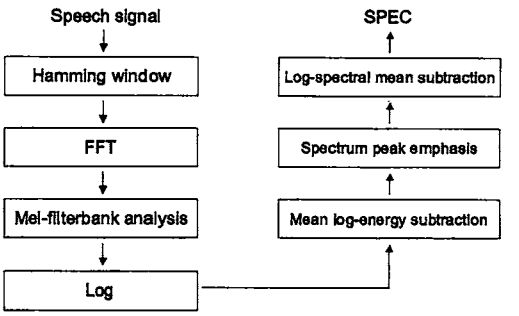

図 2 SPEC の抽出過程

ラム情報と基本周波数情報の2つのストリームを利用し，雑音に頑健な基本周波数情報のストリーム重みを大きくすることで話者照合の耐雑音性が向上することを確認している．

一方で我々は，スペクトル特徴量やケプストラム特徴量の各次元を個別のストリームとしたマルチストリーム HMM を用い，雑音環境に応じて各次元の信頼度重みの推定を行うことで，音声認識の耐雑音性が向上することも確認している [2]．先行研究 [2] では，線形判別分析（LDA）によるストリーム重み推定手法を提案しており，この手法を用いて教師ありの条件で重み推定を行っている．雑音の帯域性を効率的に抑制できることから，スペクトル特徴量の方が，ケプストラム特徴量を用いるよりも耐雑音性が高いことを確認している．

そこで本稿では，文献 [2] と同様の手法で，スペクトル特徴量である SPEC [3] と，ケプストラム特徴量である MFCC を用いてマルチストリーム HMM を構築し，それぞれを用いた話者照合手法の耐雑音性について検討を行う．[2] では特徴ベクトルの静的特徴量の次元のみを重み推定対象のストリームとしていたが，本稿では，動的特徴量を含めた全ての次元を個別のストリームとして扱い，重みの推定対象とする．また，教師なしの条件での重み推定や，評価データと異なる雑音条件のデータを重み推定に利用した，雑音条件に非依存な重みの推定を行い，その耐雑音性について検討する．

以下では，まず，本研究で使用するスペクトル領域の特徴量である SPEC について説明し，次にマルチストリームを用いた話者照合法について述べる．そして，線形判別分析に基づくストリーム重みの手法について述べ，最後に本手法の有効性を確認する話者照合実験について述べる．

2. スペクトル特徴量

本研究で使用するスペクトル特徴量 SPEC は，先行研究 [3] で提案されたもので，MFCC で行われているケプストラム領域での正規化処理と同等の処理を，スペクトル領域で行うことで作成される．文献 [3] では，音声認識において，このスペクトル特徴量 SPEC と MFCC が同程度の性能となることを確認している．図 1, 2 に，MFCC と SPEC の抽出過程を示す．どちらの手法においても，フレームごとの音声波形に対してハミング窓を用いて窓かけを行い，高速フーリエ変換を施した上で，メルフィルタバンクを用いて帯域ごとの対数スペクトルを計算する．MFCC ではこの対数スペクトルに対して離散コサイン変換を行ってケプストラム領域へ変換するが，SPEC ではその処理が行われない．そして，MFCC での C_0 項除去，リフタリング処理，ケプストラム平均除去の正規化処理の代わりに，SPEC では平均エネルギー正規化，スペクトルピーク強調，対数スペクトル平均除去の処理を行う．

3. マルチストリーム HMM による話者照合

3.1 マルチストリーム HMM

マルチストリーム HMM では，t フレーム目の入力特徴ベクトル O_t に対する出力確率の対数 $b(O_t)$ は次のように，それぞれのストリームから得られる対数尤度の重みつき線形和で計算される．

$$b(O_t) = \sum_{s=1}^{S} \lambda_s \cdot b(O_{st}) \tag{1}$$

ここで，$b(O_{st})$ はストリーム s の特徴ベクトル O_{st} の出力確率の対数であり，S は総ストリーム数，λ_s はストリーム s の重みである．

3.2 話者照合スコア

話者照合スコア $q(O)$ は，申告話者の特定話者モデル M^c から得られるフレーム平均の尤度 $p(O|M^c)$ を，不特定話者モデル M^g から得られるフレーム平均の尤度 $p(O|M^g)$ で正規化することで得られる．

$$q(O) = \log p(O|M^c) - \log p(O|M^g) \tag{2}$$

申告話者・不特定話者モデルにはマルチストリーム HMM が利用される．そこで，それぞれのモデルのストリーム s から得られる尤度 $p(O_s|M^c)$, $p(O_s|M^g)$ を用いて，この式を書き換えると，最終的に，

$$q(O) = \sum_{s=1}^{S} \lambda_s \cdot q(O_s) \tag{3}$$

となる．$q(O_s)$ はストリームごとに計算される照合スコアとなる．この照合スコアが，閾値 θ を超えた場合に，申告者本人であると判断する．したがって，判別式は $z = q(O) - \theta$ という線形関数となり，z が正であれば本人として受理，0 以下であれば詐称者として棄却する．話者照合の流れを図 3 に示す．

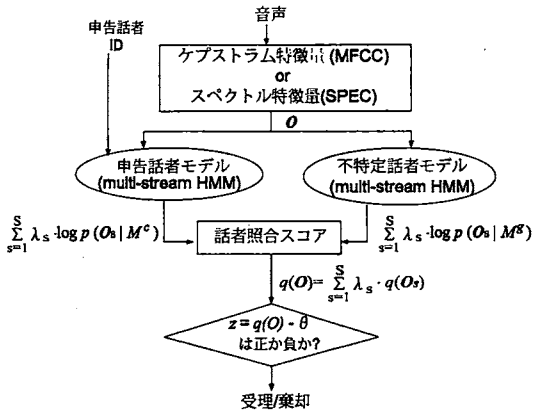

図 3 話者照合システム

4. 線形判別分析 (LDA) によるストリーム重みの推定

各ストリームから得られる照合スコアで構成される多次元空間上に，申告話者の特徴量が正しく入力されたときと，詐称者の特徴量が入力されたときのデータをプロットすることで分布を作成し，この2つの分布を識別する関数を線形判別分析 (LDA) を用いて求めると，得られる関数は，照合に用いる線形関数と同じ和の形となり，その係数をストリーム重みと見なすことが出来る．このようにすることで，重み推定用のデータの雑音条件に応じて，申告話者と詐称者の分布の識別性能が最大になるように，各ストリームの重みを推定することができる [4]．2つの分布を切り分ける模式図を図 4 に示す．

具体的には，各入力データについて，ストリーム s から得られる照合スコア $q(O_s)$ を，x_s 座標 (S 次元空間上) にプロットする．申告話者と詐称者の2つの分布について LDA を適用して得られた判別関数は，

$$a_0 + \sum_{s=1}^{S} a_s x_s = 0 \qquad (4)$$

となる．ここで，係数 a_s に負の値が算出されることがあるが，その場合にはそのストリームの信頼度が著しく低いと見なし，そのストリームの重みを 0 とする．そこで，最終的なストリーム重み λ_s は次のように計算される．

$$\lambda_s = S \cdot \frac{a'_s}{\sum_{i=1}^{S} a'_i} \quad , \quad a'_i = \begin{cases} a_i & (a_i \geqq 0) \\ 0 & (\text{otherwise}) \end{cases} \qquad (5)$$

5. 話者照合実験

まず，教師なしでストリーム重みを推定したモデルの話者照合実験を行う．次に，評価データとは異なる様々な雑音条件のデータを重み推定用データとして利用してストリーム重みを教師ありで推定することで，雑音に依存しない重みの推定が可能であるかについて検討を行う．

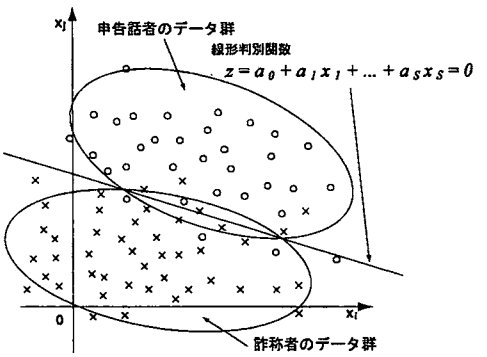

図 4 LDA の概要

5.1 実験データ

音声データには，1カ月毎に収録を行った，計 5 時期分のデータ [4] を使用した．使用した話者は男性話者 36 名分で，各話者は 1 時期に 50 個の 4 桁連続数字を発声しており，音声は 16kHz，16bit で標本化・量子化されている．

データは 12 名ずつ 3 グループに分け，それぞれを評価用，不特定話者モデルの学習用，(教師あり条件での) ストリーム重み推定用に割り当てる．詐称者としては，評価用のグループの申告話者以外の全話者を用いる．申告話者・不特定話者のマルチストリーム HMM の学習には 1 ～ 3 時期目の音声データを用い，ストリーム重みの推定と評価には，4，5 時期目のデータを用いる．

特徴量としては，ケプストラム特徴である MFCC の 25 次元ベクトル (MFCC 12 次元，ΔMFCC 12 次元，Δ 対数パワー 1 次元)，スペクトル特徴量 SPEC の 27 次元ベクトル (SPEC 13 次元，ΔSPEC 13 次元，Δ 対数パワー 1 次元) を用いた．両者は次元数は異なるが，これはスペクトル特徴量では，全帯域のエネルギーの総和が一定値になるように正規化されているためであり，情報量としては等価である．どちらの特徴量も音響分析時のフレーム周期は 10ms であり，分析窓長は 25ms である．

5.2 モデルの構築

申告話者・不特定話者モデルは数字単位のマルチストリーム HMM であり，照合スコアを得るときには，それぞれのモデルを使用して桁数制限のある連続数字認識を行う．特徴量の全ての次元を個別のストリームとするため，ストリーム数は MFCC を用いた場合は 25 ストリーム，SPEC を用いた場合は 27 ストリームである．初期モデルは，1) まず，マルチストリーム化を行っていない通常の HMM を学習し，2) 出来上がったモデルの特徴量ベクトルをストリームごと分割し，3) 全てのストリーム重みを 1.0 とすることで作成する．推定により得られたストリーム重みは，全てのモデルに対して共通に使用する．

5.3 教師なし条件でのストリーム重みの推定

教師なしの実験では，ストリーム重み推定用の話者グループは使用せず，評価用のグループの全話者のデータについて，初期モデルによる照合実験を行って，申告者か詐称者かをラベル

表 1　教師なし推定による各雑音・SN 比条件における EER (%)

		ピンクノイズ				エレベータホール雑音				列車（在来線）雑音			
		20dB	15dB	10dB	5dB	20dB	15dB	10dB	5dB	20dB	15dB	10dB	5dB
MFCC	BASE	2.9	8.4	19.8	33.1	4.6	11.3	22.6	35.5	10.0	17.6	28.7	39.5
	LDA	2.7	7.6	20.1	35.0	3.5	10.0	22.5	36.3	7.4	16.9	29.4	41.0
SPEC	BASE	2.5	7.3	18.3	34.2	5.0	12.1	23.2	36.8	9.8	16.7	27.5	39.4
	LDA	2.1	6.8	19.5	35.6	3.5	10.6	23.7	36.9	7.0	15.7	27.9	40.3

表 2　雑音種ごとの平均誤り削減率 (%)

特徴量	ピンクノイズ	エレベータホール雑音	列車 (在来線) 雑音
MFCC	2.4	8.5	5.9
SPEC	2.7	9.6	7.8

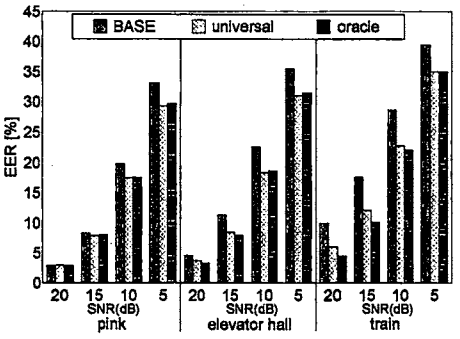

図 5　MFCC における各雑音・SN 比での EER

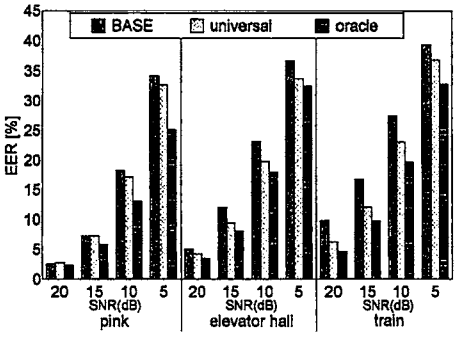

図 6　SPEC における各雑音・SN 比での EER

付けし，その判定結果をもとにして重み推定を行う．本研究では，初期モデルを用いた照合の際の閾値は，等誤り率（EER）が最小となるように定めた．6 通りの話者グループの組み合わせで照合実験を行い，全ての結果を平均して全体の評価に用いた．

学習用データには SN 比 30 dB でピンクノイズを，評価用データには，ピンクノイズ，電子協騒音データベース [5] 中の，エレベータホール，列車 (在来線) 雑音を SN 比 5, 10, 15, 20 dB で付加した．

各雑音・SN 比条件において，MFCC, SPEC それぞれについて初期モデル (BASE) の EER と教師なし重み推定後のモデル (LDA) の EER を表 1 に示す．また，初期モデルからの EER の平均誤り削減率を，雑音種ごとに各 SN 比から得られた結果を平均して表 2 に示す．

表 1 を見ると MFCC, SPEC どちらの特徴量を用いても，SN 比が 15, 20 dB では全ての雑音において初期モデルと比べて誤り率が削減されていることが分かる．ストリーム重み推定の効果は，初期モデルによる照合誤りが大きくなると得られにくくなり，等誤り率がおおよそ 20% 以上になると改善効果が得られなくなる．また，表 1 では MFCC と SPEC のベースライン (BASE) の性能の優劣は見られないが，表 2 を見るとスペクトル特徴量である SPEC の方が，MFCC よりも改善が大きい傾向があることが分かる．これは，我々の音声認識の先行研究 [2] で得られた「帯域性のある雑音の影響を抑制する効果が SPEC の方が大きい」という結果と同様の傾向である．

5.4　雑音環境に非依存なストリーム重みの推定

雑音に非依存なストリーム重みを推定するために評価データに重畳する，ピンクノイズ，エレベータホール雑音，列車 (在来線) 雑音の他に，電子協騒音データベースの走行車内，駅，計算機室雑音も用意する．重み推定用データに評価データと異なる 5 種類の雑音を，SN 比 5, 10, 15, 20 dB で重畳し，これら全てを用いて教師ありで重み推定を行って評価データの雑音条件に依存しない重みの推定を行う．また，比較のために，評価データと同じ雑音，SN 比条件で重み推定用データに雑音を重畳して重み推定を行った場合（雑音条件が既知とした場合）についての照合実験も行った．

各雑音・SN 比条件において，MFCC, SPEC それぞれについて初期モデル (BASE), 雑音環境に非依存な重み推定をした場合 (universal), 雑音環境を既知として重み推定を行った場合 (oracle) の EER を図 5, 6 に示す．全ての雑音環境において，初期モデルからの誤り率の削減が得られ，評価データの雑音環境に非依存な重みが推定されていることが分かる．また，MFCC では雑音に非依存な重み推定をした場合と，雑音を既知として重み推定をした場合の結果にほとんど差は見られなかった．一方 SPEC では，雑音環境を既知として重み推定を行った場合には大きな性能改善効果が得られているのに対して，雑音環境に非依存な重み推定ではその効果が小さかった．これは，MFCC に比べて SPEC の方が帯域性雑音の特徴が反映されやすいため，雑音種が特定された条件下であれば，SPEC に対して重み付けを行った方が雑音抑制の効果が出やすいが，評価データと異なる雑音種を用いて推定を行った場合には，その効果が得られにくいことを示している．

5.5　推定された重みの比較

MFCC と SPEC それぞれについて，列車 (在来線) 雑音の

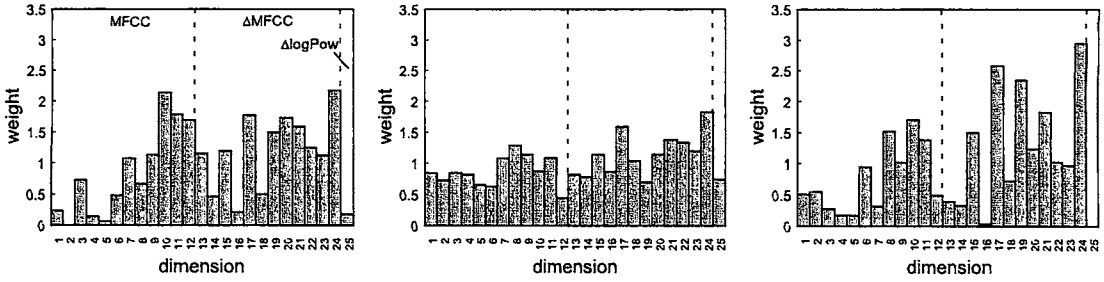

図 7　MFCC での列車 (在来線) 雑音の SN 比 15dB における重み．左図: 雑音条件既知で教師ありで推定した重み，中央図：教師なしで推定した重み，右図：雑音条件に非依存な重み

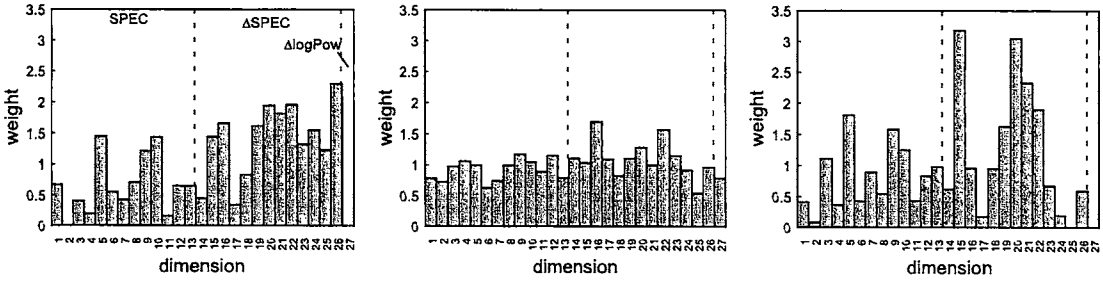

図 8　SPEC での列車 (在来線) 雑音の SN 比 15dB における重み．左図：雑音条件既知で教師ありで推定した重み，中央図：教師なしで推定した重み，右図：雑音条件に非依存な重み

SN 比 15dB における，雑音条件を既知とした教師ありで推定した重みと教師なし条件で推定した重み，雑音条件に非依存に推定した重みを図 7, 8 に示す．重みは，6 通りの実験の結果の平均である．図 7 では，1〜12 次元が MFCC，13〜24 次元が ΔMFCC，25 次元が Δ 対数パワーであり，図 8 では，1〜13 次元が SPEC，14〜26 次元が ΔSPEC，27 次元が Δ 対数パワーである．図 7, 8 を見ると，どちらの特徴量においても，全体として静的特徴量の部分よりも動的特徴量に重みが付与されている．また，教師なし条件で推定した重みを，教師あり条件で推定した重みと比べると，MFCC, SPEC ともに，重要な次元に対して重み付けがなされず，全体的に平坦になっていることが分かる．また，雑音条件に非依存な重みを，雑音条件既知で推定した重みと比べると，MFCC では両者は大局的に似た傾向となっているが，SPEC では重みの傾向が異なっている．これは，SPEC の方が重み推定に用いる雑音種の影響を受けやすいためと考えられる．

6. ま と め

本稿では，ケプストラム特徴量 (MFCC) と スペクトル特徴量 (SPEC) を用いて，特徴量の各次元をストリームして扱ったマルチストリーム話者照合の耐雑音性について検討を行った．各ストリームの重みは線形判別分析 (LDA) を用いた手法により推定し，4 桁連続数字を用いた照合実験において本手法の有効性を確認した．実験の結果，教師なし条件でストリーム重みを推定することによって，様々な雑音条件の SN 比 15, 20 dB において，初期モデルからの誤りが削減され，耐雑音性の向上が確認された．

また，複数の雑音条件の重み推定用データを用いることによって，評価データの雑音環境に依存しないストリーム重みが推定され，誤り率が削減することを確認した．

今後の課題として，より多くの音声データを用いた場合の本手法の有効性の確認が必要である．これは，話者グループの組合せによって推定されたストリーム重みにばらつきがあったためである．他の課題として，より多くの雑音種を用いた場合の有効性の確認や，他のストリーム重みの推定法の導入などが挙げられる．

7. 謝　　辞

本研究の一部は文部科学省科学研究費補助金若手研究 (B) No.17700141 の支援を受けて実施された．

文　　献

[1] 浅見太一，岩野公司，古井貞熙，“雑音に頑健な話者照合のための基本周波数情報の利用，” 信学技報，vol.104，no.87，pp.1-6 (2004-5)

[2] 岩野公司，小島要，古井貞熙，“マルチバンド音声認識のための LDA に基づく帯域重み推定手法，” 信学技報，SP2006-3，pp.65-66 (2006-5)．

[3] 西村義隆，篠崎隆宏，岩野公司，古井貞熙，“周波数帯域ごとの重みつき尤度を用いた雑音に頑健な音声認識，” 信学技報，vol.103，no.519，pp.19-24 (2003-12)

[4] 浅見太一，岩野公司，古井貞熙，“マルチストリーム話者照合におけるブースティングに基づく閾値最適化法の検討，” 信学技報，SP2005-90，pp.1-6 (2005-12)

[5] S. Itahashi, “Recent speech database projects in Japan,” *Proc. ICSLP 1990*, vol.2, pp.1081-1084, Kobe, Japan, November 1990.