区分的線形変換を用いた雑音環境下マルチモーダル音声認識*

☆柏木陽佑, 鈴木雅之, 峯松信明, 広瀬啓吉 (東大)

1 はじめに

近年，インターフェースとしての利便性が期待され，音声認識技術が非常に注目されている．統計的手法の確立などによってクリーンな音声に対する認識率は向上しているが，高雑音環境下では認識率が著しく低下してしまう．この問題を解決するための手法の一つとして，特徴量強調の研究がなされている．例えば Stereo-based Piecewise Linear Compensation for Environments (SPLICE) [1] や Vector Taylor series(VTS) [2] を用いる手法などがある．特徴量強調の他にも，雑音に対してモデルを適応する手法，特徴量をあらかじめなんらかの方法で正規化する手法などがある [3].

それらとは異なるアプローチとして，音声と同時に様々な情報を利用することによって認識精度を向上させるマルチモーダル手法がある．これは顔の表情や口唇の動きなどと発話には相関があると考え，それらを認識の際に利用するものである．また，認識に直接利用するだけでなく音声区間検出に用いる例などもある [4] [5]．マルチモーダル音声認識は大きく分けて音声と画像別々に認識して得られた尤度を統合する結果統合 [6–9] と特徴量ベースで結合して認識を行う初期統合 [9] [10] の 2 種類がある．結果統合は音声のモデルと画像のモデルそれぞれから得られる尤度を重みづけにより統合する．結果統合では，デコーディングを複数回行ない，それらを重みつけする必要がある．それに対し，初期統合は特徴量を変更するだけで，モデルの学習が比較的簡単であるという特徴がある．したがって，本稿ではこの初期統合に注目する．

初期統合の問題点は特徴量をどのような基準で統合するかである．この基準は，雑音の種類や大きさ，発話の音素の種類などによって適切に変更するべきであると考えられる．本稿では，これを実現するために，区分的線形変換を用いた新しい初期統合法を提案する．これは，劣化音声特徴量をガウス混合分布（Gaussian Mixture Model : GMM）を用いることで多くのクラスに分類し，それぞれのクラスに所属する音声特徴量と画像特徴量の連結ベクトルに対する最適な線形変換を推定する．これによって，クラスに依存した最適な基準で音声と画像の特徴量を利用することができる．本手法は劣化音声特徴量からクリーン音声特徴量を区分的線形変換によって推定する SPLICE を参考としている．SPLICE との違いは，画像情報も利用して区分的線形変換を行っている点である．

2 章でマルチモーダル音声認識の要素技術ついて述べる．その後，3 章で提案手法，4 章で実験について述べる．最後に 5 章でまとめについて述べる．

2 マルチモーダル音声認識

雑音環境では音声特徴量が大きく歪むため，音声のみによる音声認識では認識率が低下してしまう．この問題は，音響的なノイズに独立な情報を用いることで改善することができる．マルチモーダル音声認識は音声と顔の画像（表情や口唇の動き）には相関があると考え，それらを両方利用することで，ノイズにロバストな音声認識システムを構築することを目標とする．マルチモーダル音声認識において，どのような画像特徴量が発話を適切に表現しているか，そして画像特徴量をどのように認識プロセスに組み込むかが問題となる．

2.1 画像特徴量

用いられる画像特徴量は apeearance ベース，shape ベース，そしてその両方を用いたものの大きく 3 種類に分かれる．appearance ベースは画像の全てのピクセルが発声と相関があると考え，対象領域内のピクセル値を用いる．しかし，そのままではピクセル数分の次元数を持ってしまうため，特徴量の次元を減らすためにさらに次元圧縮をしたものを用いることが多い．それに対し，shape ベースは口唇やその他の顔のパーツの型値が発声に相関を持っていると考え，例えば口唇の高さや幅などから特徴量を抽出する．これは比較的照明の変化などに強いという特徴があるが，口唇の輪郭を正確に抽出する必要があるために，計算量が高くなるなどの問題がある．両方を用いた抽出法では，それぞれで得られた特徴量を結合し，Active Appearance Model [12] のようなモデルを用いて学習する．本論文では画像のノイズは想定せず，音声認識における新しい初期統合法の提案に主眼を置いているため，eigen-face を用いた appearance ベースの特徴量を用いた [13].

2.2 音声-画像の統合

マルチモーダル音声認識において，音声と画像の統合方法には大きく分けて初期統合（feature fusion）と結果統合（decision fusion）の 2 種類ある．初期統合は特徴量の段階で統合するものであり，音声特徴量と画像特徴量を連結し，この連結特徴量を HMM などを用いてモデル化する．また，単にそのまま使うだけでなく更に PCA(principle component analysis) などをかけて次元圧縮したベクトルを利用することもできる．それに対して，結果統合はマルチストリーム HMM を用いる．各ストリームで認識を行い，音声 HMM から得られた対数尤度を L_a，画像 HMM か

*Multi-modal automatic speech recognition using piecewise linear transformation in noisy environments by Y.Kashiwagi, M.Suzuki, N.Minematsu, and K.Hirose (The University of Tokyo)

ら得られた対数尤度を L_v とすると，それらを式 (1) のように統合する．

$$L_{av} = \lambda_a L_a + \lambda_v L_v$$
$$\text{where } \lambda_a + \lambda_v = 1, \lambda_a, \lambda_v \geq 0 \quad (1)$$

λ_a，λ_v はそれぞれ音声モデルと画像モデルから出力される対数尤度に対する重みである．しかし，対数尤度で統合するため，各ストリームでアライメントが一致しない可能性がある．この問題を軽減するために，音声 HMM から得られた強制アライメントを用いて画像 HMM を構築する．

3 提案手法

前述のように結果統合法は各ストリームごとに認識を行い，対数尤度で統合するためにそれぞれのストリームにおける特徴量のアライメントが取れる保証はない．それに対し，初期統合は特徴量の段階において統合し，統合した特徴量をモデル化して認識する．そのため，アライメントの問題はなくなるが，このままでは次元数が大きくなってしまう．次元圧縮手法として PCA などがあり，単一のノイズに対しては非常に効果的である．しかし，複数種類のノイズを想定した場合，ノイズを推定して適応的に変化させなければ，全てのノイズに対して同一の変換をしてしまうために，ロバスト性が低下してしまう．

本提案手法では初期統合におけるこの問題を解決するため，観測音声信号に依存させて音声と画像の情報を適切に組み合わせる新しい音声と画像情報の統合法を提案する．画像特徴量 $\boldsymbol{i}$ の系列と音声特徴量 $\boldsymbol{y}$ の系列が同期していることを仮定し，入力として連結ベクトル，出力として音声特徴量の推定値 $\hat{\boldsymbol{x}}$ の区分的線形変換を式 (2) のようにする．

$$\hat{\boldsymbol{x}} = \sum_k P(k|\boldsymbol{y}) \boldsymbol{A}_k \boldsymbol{m}' \quad (2)$$

ただし，$\boldsymbol{m}' = [1, \boldsymbol{y}^\top, \boldsymbol{i}^\top]^\top$ である．画像特徴量 $\boldsymbol{i}$ は eigen-face を用いた appearance ベースを利用する．$\boldsymbol{y}$ の確率密度関数を GMM と仮定すると $P(k|\boldsymbol{y})$ は

$$\begin{aligned} P(\boldsymbol{y}|k) &= \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ P(\boldsymbol{y}) &= \sum_k \pi_k \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ P(k|\boldsymbol{y}) &= \frac{\pi_k \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \end{aligned} \quad (3)$$

となる．ただし，GMM の分散 $\boldsymbol{\Sigma}_k$ は対角行列として学習している．その後，線形変換 $\boldsymbol{A}_k$ を連結ベクトル $\boldsymbol{m}'$ とクリーン音声 $\boldsymbol{x}$ のパラレルデータを用いて重み付き最少二乗誤差基準で学習する．

$$\{\boldsymbol{A}_k\} = \underset{\{\bar{\boldsymbol{A}}_k\}}{\operatorname{argmin}} \sum_l \sum_k P(k|\boldsymbol{y}) \left\| \boldsymbol{x}_l - \bar{\boldsymbol{A}}_k \boldsymbol{m}' \right\|^2 \quad (4)$$

これは解析的に解くことができ，

$$\hat{\boldsymbol{A}}_k = \boldsymbol{X} \boldsymbol{R}_k \boldsymbol{M}'^\top (\boldsymbol{M}' \boldsymbol{R}_k \boldsymbol{M}'^\top)^{-1} \quad (5)$$

となる．ここで，$\boldsymbol{M}'$ は $\boldsymbol{m}'$ を並べたもの，$\boldsymbol{R}$ は対角成分に $[P(k|\boldsymbol{y}_1), P(k|\boldsymbol{y}_2), \ldots, P(k|\boldsymbol{y}_L)]$ を持つ対角行列となる．本手法は音声と画像の連結ベクトル $\boldsymbol{m}$ から音声特徴量に変換する線形変換 $\boldsymbol{A}_k \boldsymbol{m}'$ を，$\boldsymbol{y}$ を観測した際の k の事後確率 $P(k|\boldsymbol{y})$ を重みとして足し合わせることで音声強調を実現している．すなわち，$\boldsymbol{A}_k \boldsymbol{m}'$ の計算が，特徴量を次元圧縮していることに相当している．一方，事後確率 $P(k|\boldsymbol{y})$ によって，入力音声に依存して利用した線形変換の選択を重み付けによって実現する．これによって，式 (1) は入力に依存させて適切な次元圧縮を選択することとなり，精度の良い雑音抑圧が期待される．

3.1 SPLICE との比較

SPLICE は特徴量強調手法の一つである．劣化音声特徴量 $\boldsymbol{y}$ からクリーン音声特徴量 $\boldsymbol{x}$ を区分的線形変換を用いて推定する．

$$\hat{\boldsymbol{x}} = \sum_k P(k|\boldsymbol{y}) \boldsymbol{A}_k \boldsymbol{y}' \quad (6)$$

$\hat{\boldsymbol{x}}$ は推定されたクリーン音声特徴量，$\boldsymbol{y}$ は観測音声特徴量であり，$\boldsymbol{y}' = [1, \boldsymbol{y}^\top]^\top$ である．$\boldsymbol{y}$ の確率密度関数は GMM で表すことができると仮定する．本提案手法は SPLICE と非常に良くにており，違いは線形変換部分である．SPLICE は入力に観測音声特徴量のみを用いるが，提案手法は音声特徴量と画像特徴量の連結ベクトルを用いてクリーン音声特徴量を推定する．

4 実験

提案手法の有効性を示すため，CENSREC-1-AV [14] と noisex92 [15] を用いて実験を行った．用いたデータを表 1，表 2 に示す．クリーン音声データは CENSREC-1-AV のものを用い，ノイジー音声データは noisex92 の雑音データを重畳することによって作成した．SNR は 20dB から 0dB までの 5 段階を採用した．1 発話はそれぞれ 1 〜 7 桁の日本語数字読み上げから構成されており，音声データと画像データは同期が取れているものとする．画像に関してはクリーン条件でのみ学習・認識を行った．表 3 に実験に用いた特徴量を示す．音声特徴量は MFCC13 次元とそれの一次微分，二次微分を用いる．画像特徴量はカラー画像を用いたため，Raster scan をしてベクトル化した後に，それを PCA により次元圧縮して得られた各色 10 次元の計 30 次元を用いている．その後，音声と画像のフレームレートが違うため，画像特徴量を線形変換して音声とフレーム単位で同期させた．

認識用 HMM 学習データセット，線形変換学習用パラレルデータセット，テストデータセットは表 4 のようにした．認識用 HMM はクリーン音声でのみ学習したもの（clean HMM）とクリーン音声だけでなく 5 段階の SNR で雑音を重畳した劣化音声も用いて学習した（multi-condition HMM）2 パターンを作成した．また，前述の様に線形変換の学習にはノイジー

Table 1　Audio data.

Samplling rate	16kHz
Auantization bits	16 bit/sample
Audio noise	noisyx-92 5 types (babble,factory1, factory2,car (Volvo),white) 5 SNR levels (20 dB to 0dB)

Table 2　Visual data.

Frame rate	29.97Hz
Pixsel	24 bit color
Data size	81 pixel width × 55 pixel height
Visual noise	none (clean)

Table 3　Audio and visual features.

Audio	MFCC+Δ+Δ^2
Visual	Raster scan + PCA (10 × 3 (RGB) = 30 dimensions)

Table 4　Data sets for training and testing.

	HMM	Trans.	Test
Speaker	male 22 female 20		male 25 female 26
Audio data	clean noisyx-92 (babble, factory1, factory2, car (Volvo), white) 20dB 15dB 10dB 5dB 0dB		
visual data	clean color (RGB)		

音声とクリーン音声のパラレルデータが必要だったため，CENSREC-1-AV の学習データにノイズを重畳することによって擬似的なパラレルデータを得た．提案手法と従来手法を認識実験により比較する．従来手法には，音声強調をかけないもの (no enhance)，初期統合 (feature fusion (PCA))，結果統合 (decision fusion) を用いた．初期統合は音声と画像の特徴量を連結した 69 次元のベクトルを PCA により 39 次元に次元圧縮を行っている．結果統合は，式 (1) における対数尤度を統合する際の重み λ を 0.0，0.2，0.4，0.6，0.8，1.0 の内から最も良い認識率が得られるものを各雑音，各 SNR ごとに選択している．結果統合は音声と画像の各ストリームでアライメントが一致しない問題を軽減するために，画像 HMM を音声 HMM より得られるアライメントを用いて構築した．また，画像を用いることによる効果を調べるために，音声特徴量のみを用いる SPLICE とも比較をおこなった．全ての場合において，CMN(Cepstral Mean Normalization) をかけている．

音声強調した後，認識実験をした結果を Fig. 1 に示す．これは，全ての雑音と SNR における単語誤り率を平均したものである．clean 条件と multi 条件は，それぞれクリーン音声のみで学習した HMM と雑音も用いて学習した HMM による認識結果である．また，全ての場合において，HMM は 39 次元の特徴量ベクトルで学習を行った．実験の結果，従来手法 (no enhancement, feature fusion, decision fusion) の中では最も結果統合が良い結果を得られた．提案手法は，結果統合よりもクリーン音声のみで学習した HMM で 25%，雑音も用いて学習した HMM で 24%のエラー削減率を得ることができた．また，SPLICE と比較した場合も提案手法が認識率が良い結果となった．これにより，区分的線形変換に画像特徴量も用いることの効果を示すことができた．

Fig. 2–6 はそれぞれの種類の雑音ごとに全ての SNR の結果を平均したものである．ほぼ全ての条件で提案手法が最も良い結果を得ることができた．例外として，走行時雑音 (car noise) における multi 条件では，結果統合が最も良い結果を示したが，これは，そもそも何も耐雑音処理をしない場合 (no enhancement) でも非常に単語誤り率が低いことが原因である．全ての雑音の SNR においても，もともと単語誤り率が非常に低い箇所では同様の傾向が見られる．

5　まとめ

本稿では，区分的線形変換に基づくマルチモーダル音声認識手法を提案した．これは，劣化音声特徴量と口唇画像の特徴量の連結ベクトルとクリーン音声特徴量のパラレルデータを用いることによって学習を行い，入力音声によって柔軟に変換を変える枠組みである．CENSREC-1-AV を用いた認識実験により，提案手法は，結果統合よりもクリーン音声のみで学習した HMM で 25%，雑音も用いて学習した HMM で 24%のエラー削減率を得ることができた．また，雑音音声からクリーン音声を区分的線形変換によって推定する SPLICE と比較した場合でも本提案手法は良い結果を得ることができた．これによって，実験により特徴量強調に画像情報を利用することの有効性が示された．

参考文献

[1] J. Droppo, *et.al.*, “Evaluation of the SPLICE on the Aurora2 and 3 Tasks,” Proc. ICSLP, pp. 29–32, 2002.

[2] V. Stouten, “Robust Automatic Speech Recognition in Time-varying Environments,” PhD thesis, 2006.

[3] J. Droppo, *et. al.*, “Environmental Robustness,” Benesty, Sondhi, Huang (eds) Handbook of Speech Processing, pp. 653–679, 2008.

[4] S. Tamura, *et.al.*, “A Robust Audio-Visual Speech Recognition Using Audio-Visual Voice Activity Detection,” Proc. Interspeech, 2010.

[5] I. Almajai and B. Milner, “Using audio-visual features for robust voice activity detection in clean and noisy speech,” Proc. EUSIPCO, 2008.

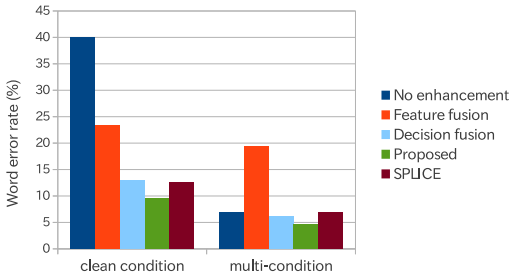

Fig. 1 Word error rates for clean HMMs and multi-condition HMMs.
The results are averaged over all noise types and levels.

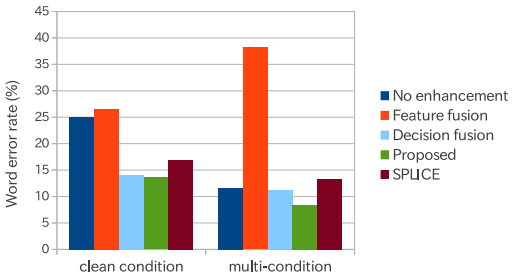

Fig. 2 Word error rates for clean HMMs and multi-condition HMMs in babble noise condition.
The results are averaged over all noise levels.

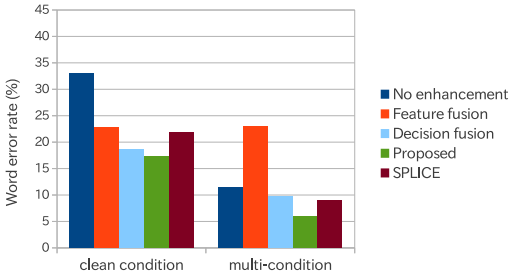

Fig. 3 Word error rates for clean HMMs and multi-condition HMMs in factory1 noise.
The results are averaged over all noise levels.

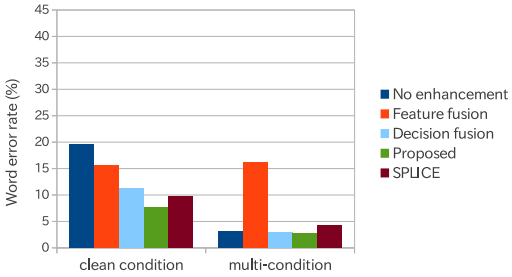

Fig. 4 Word error rates for clean HMMs and multi-condition HMMs in factory2 noise.
The results are averaged over all noise levels.

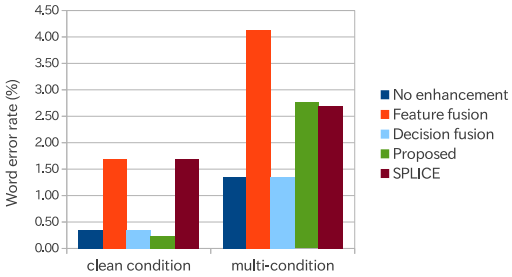

Fig. 5 Word error rates for clean HMMs and multi-condition HMMs in car noise (Volvo noise).
The results are averaged over all noise levels.

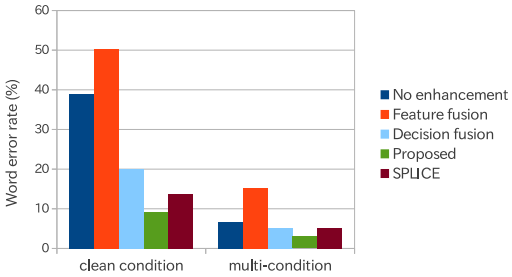

Fig. 6 Word error rates for clean HMMs and multi-condition HMMs in white noise.
The results are averaged over all noise levels.

[6] S. Dupont and J. Luettin, "Audio-visual speech modeling for continuous speech recognition," IEEE Trans. Multimedia, vol. 2, pp. 141–151, 2000.

[7] G. Potamianos and H. P. Graf, "Discriminative training of HMM stream exponents for audio-visual speech recognition," Proc. ICASSP, pp. 3733–3736, 1998.

[8] M. Tariquzzaman, *et. al.*, "Performance Improvement of Audio-Visual Speech Recognition with Optimal Reliability Fusio," ICICIS, pp. 203–206, 2011

[9] G. Potamianos, *et. al.*, "Recent advances in the automatic recognition of audiovisual speech," Proc. IEEE, pp. 1306–1326, 2003.

[10] G. Potamianos, *et. al.*, "Hierarchical discriminant features for audio-visual LVCSR," Proc. ICASSP, pp. 165–168, 2001.

[11] Kunikoshi, *et. al.*, "Speech generation from hand gestures based on space mapping," Proc. INTERSPEECH, pp. 308–311, 2009.

[12] Cootes, T.F., "Active Appearance Model," Proc. European Conference on Computer Vision, vol. 2, pp. 484–498, 1998.

[13] Turk, M.A., *et. al.*, "Face recognition using eigenfaces," CVPR, pp. 586–591, 1991.

[14] S. Tamura, *et al.*, "CENSREC-1-AV: An audio-visual corpus for noisy bimodal speech recognition," Proc. AVSP, 2010.

[15] Varga, A. and Steeneken, H.J.M., "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," Speech Communication, Vol. 12, pp. 247–251, 1993.