音響尤度のリスコアリングによる結果統合を用いたバイモーダル連続音声認識 *

◎石川 剛　南角 吉彦　全 炳河　宮島 千代美　徳田 恵一　北村 正　(名工大)

1. はじめに

バイモーダル音声認識では，音声情報と唇動画像情報を統合する手法として，初期統合や結果統合，合成統合 [1] が提案されている．合成統合は，音声と画像の相関をうまく利用しながら音声と画像の非同期性も表現できるモデルであるが，音素境界を越えたずれを表現することができない．このような問題に対し，音素境界にずれを表現する状態を追加学習することによる時間のずれを考慮したモデルが提案されている [2]．本研究では，初期統合により生成したラティスを利用して音響尤度のリスコアリングを行うことにより，音声と画像の非同期を考慮した認識を容易に実現する手法を提案する．

2. 音声と画像の統合

バイモーダル音声認識における音声と画像の統合方法には，時間的な統合単位の違いから，初期統合や結果統合などが提案されている．結果統合は，音声と画像を個別に認識しそれらの尤度を統合する手法である．結果統合において，統合された最終的な尤度は次式により計算することができる．

$$P(\boldsymbol{O}|M) = P(\boldsymbol{O}_{\mathrm{A}}|M_{\mathrm{A}})^{\lambda_{\mathrm{A}}} \times P(\boldsymbol{O}_{\mathrm{V}}|M_{\mathrm{V}})^{\lambda_{\mathrm{V}}} \quad (1)$$

ここで，$P(\boldsymbol{O}|M)$ は統合後の尤度，$P(\boldsymbol{O}_{\mathrm{A}}|M_{\mathrm{A}})$, $P(\boldsymbol{O}_{\mathrm{V}}|M_{\mathrm{V}})$ は音声と画像それぞれの尤度である．また，λ_{A}, λ_{V} は音声と画像それぞれのストリーム重みである．結果統合では非同期性を表現することができるが，音声と画像の相関を考慮しないため，音声と画像が音素レベルで相補的であるという特徴が有効に利用されていない．一方，初期統合はフレーム単位で特徴量を統合する手法である．初期統合における，状態 i における出力確率は次式で表される．

$$b_i(\boldsymbol{o}_t|M) = b_{\mathrm{A}i}(\boldsymbol{o}_{\mathrm{A}t}|M)^{\lambda_{\mathrm{A}}} \times b_{\mathrm{V}i}(\boldsymbol{o}_{\mathrm{V}t}|M)^{\lambda_{\mathrm{V}}} \quad (2)$$

ここで，$b_{\mathrm{A}i}(\boldsymbol{o}_{\mathrm{A}t}|M), b_{\mathrm{V}i}(\boldsymbol{o}_{\mathrm{V}t}|M)$ は音声および画像の時刻 t，状態 i における確率である．初期統合は音声と画像の同期性を前提としており 2 つの情報の相補性を利用することができる．しかし，実際には音声と画像の動きは多様な時間ずれを含んでおり，これらを学習するためには学習データが不十分な場合が多い．また，初期統合と結果統合の中間的手法に product HMM を用いた合成統合がある．product HMM は音声と画像の HMM の直積で定義され，音声と画像の時間ずれの関係を学習することができる．

3. リスコアリングを用いた統合法

product HMM は音声と画像の相関を考慮しつつ，非同期性を表現することができ，音声と画像が音素レベルで相補的であるという特徴が有効に利用できる．しかし，連続音声認識では，product HMM を 1 本のパスで連結するため，モデル境界における音声と画像のずれを表現することができない．このような問題に対し，音素境界にずれを表現する状態を追加学習することによる時間のずれを考慮したモデルが提案されている [2]．しかし，この手法は HMM の再学習や従来のデコーダを改良する必要があり計算が複雑になる．そこで，本研究では，初期統合と結果統合を活用することにより，容易に音素や単語間のずれを許容した形で認識する手法を提案する．提案法では，まず，初期統合法によるバイモーダル連続音声認識を行いラティスを生成する．次に，得られたラティスの音響尤度を，音声と画像で個別に学習されたモデルによりそれぞれリスコアリングし，ラティス中で最も尤度の高いパスを結果とする．図 1 に本システムの概要を示す．図のように，提案法は音声と画像の個別のモデルによりリスコアリングするため，音素境界を越えた時間のずれを表現することができる．また，初期統合法によりラティスを生成するため，音声と画像の相関を考慮したパスが得られると考えられる．

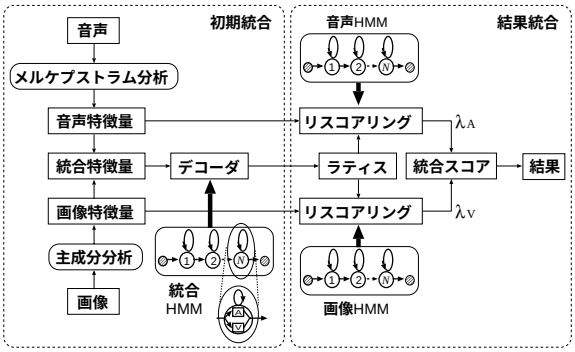

図 1．認識システム

4. 認識実験

4.1 実験条件

本実験では M2TINIT データベース [3] を使用する．このデータベースは男性話者 1 名の唇動画像とその音声が収録されている．発話内容は ATR 日本語音声データベースの音韻バランス文 503 文章で用いられているテキストである．学習データには 450 文章を使用し，残りの 53 文章をテストデータとして認識実験を行った．

音声の特徴量にはメルケプストラム係数，画像の特徴量には PCA による主成分スコアを用いた．音声と画像の分析条件の詳細は表 1 の通りである．初期統合では音声と画像データのフレーム周期を同期させるため，画像フレームをコピーして補間した．PCA は，学習データからランダムに選んだ 1000 枚の画像に対して行った．予備実験の結果に基づき，初期

∗ Bimodal Continuous Speech Recognition Using Late Integration Method Based on Acoustic Likelihood Rescoring.
By Tsuyoshi Ishikawa, Heiga Zen, Yoshihiko Nankaku, Chiyomi Miyajima, Keiichi Tokuda, and Tadashi Kitamura (Nagoya Institute of Technology)

統合における統合モデルは，状態数 3，混合数 16 の triphone モデルとした．また，リスコアリング時に用いる音声モデルと画像モデルの混合数は 8 とした．音声の重みは結果統合，初期統合ともに $\lambda_A + \lambda_V = 2$ として実験を行った．言語モデルには，IPA 日本語ディクテーション基本ソフトウェア付属の毎日新聞 45ヶ月分によって作成された 2 万語彙の言語モデル (bigram) を用いた．

表 1．音声と画像の分析条件

音声	分析窓 : Blackman 窓, 分析窓長 : 25ms 特徴量 : メルケプストラム (18 次), Δ, Δ^2 フレーム周期 : 10ms
画像	特徴量 : 主成分スコア (15 次元), Δ, Δ^2 フレーム周期 : 33.3ms $\rightarrow$ 10ms

4.2 実験結果

テストデータとして clean な音声と SN 比が 24dB, 12dB, 6dB となるように音声にガウス雑音を加えた音声を用意した．初期統合でラティスを生成する際の音声重みは，各 SN 比に対し予備実験において最もよい結果となった値で固定した．また，言語重みと挿入ペナルティについても，各 SN 比ごとに設定した．初期統合による単語正解精度と生成した 100 ベスト候補より得られる単語正解精度の上限を表 2 に示す．これより 100 ベスト候補でもリスコアリングにより認識率が改善される余地があることがわかる．

次に，初期統合で得られた N ベスト候補を音声のみと画像のみでアライメントを取り直した場合の，音素境界における画像に対する音声のずれを図 2 に示す．フレーム数が正のときに，音声が画像よりも遅れていることを表す．図より，フレームのずれはおよそ 2 フレーム程度に収まっていることがわかる．

続いて，提案法に関して認識実験を行った．まず，はじめに，リスコアリングを行う際に，アライメントに特に制約を設けない場合について実験を行った．次に，単語境界や音素境界を初期統合で得られた時刻に固定してアライメントをとるという制約を設けた場合について実験を行い，さらに，初期統合時の境界から数フレームのずれを許して探索を行う場合についての実験を行った．ただし，リスコアリング時には，挿入ペナルティや言語重みは初期統合と同じものを用い，音響尤度のみを更新した．また，音声の重みは，新たに各 SN 比，各制約下において最もよい結果となる値を用いた．単語境界，音素境界を制約した場合の実験結果をそれぞれ図 3，4 に示す．結果より，clean では初期統合の認識率を上回らなかったが，雑音状況下においては初期統合のみによる認識率よりも改善されていることがわかる．また，雑音の影響が大きい 12dB, 6dB においては音素境界におけるフレームのずれを制限することで，より大きい改善が得られた．これは音声の雑音が強い場合に音声のアライメントが大きくずれるためであると考えられる．

5. むすび

本研究では，初期統合と結果統合を利用することにより，音声と画像の相関と非同期性を考慮した認識手法を提案した．また，認識実験により，簡単な実装で初期統合の結果を改善できることを確認した．今後の課題として，初期統合からラティスを生成するための候補を増やした場合における提案法の有効性の検討などが挙げられる．出現頻度 (音素数)

表 2．初期統合の結果と認識率の上限

	clean	24dB	12dB	6dB
音声のみ	73.4%	49.7%	8.7%	0.5%
初期統合 (音声重み)	74.7% (1.6)	66.1% (1.0)	37.8% (0.8)	25.5% (0.4)
100 ベストの上限	83.7%	79.4%	57.2%	40.4%

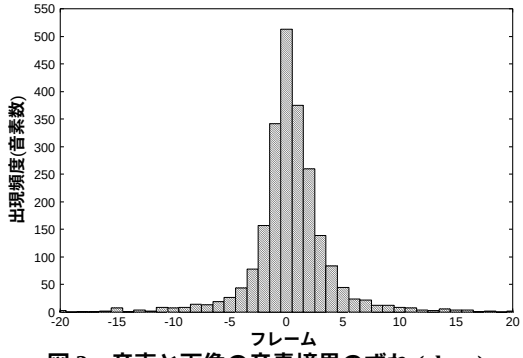

図 2．音声と画像の音素境界のずれ (clean)

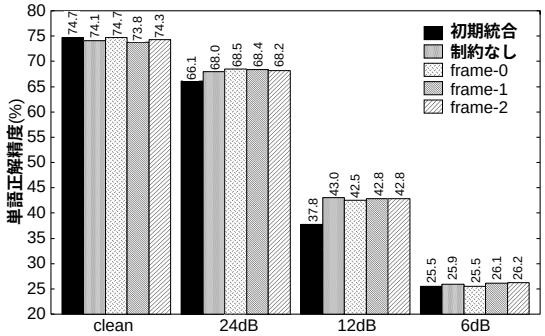

図 3．単語境界制約による単語正解精度結果

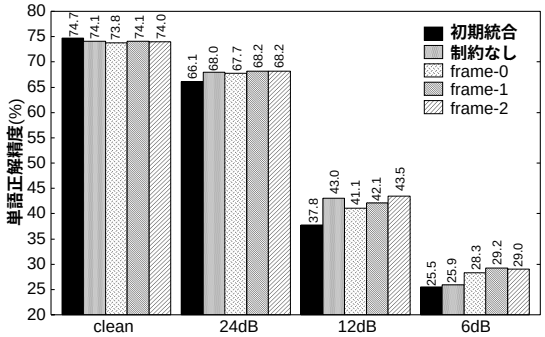

図 4．音素境界制約による単語正解精度結果

参考文献

[1] J. Luettin, G. Potamianos, and C. Neti, “Ansynchronous stream modeling for large vocabulary audio-visual speech recognition,” *Proc. ICASSP 2001*, vol.1, pp.169–172, May 2001.

[2] 中村哲，熊谷健一，田村哲嗣，“バイモーダル音声認識における音素境界を越えた同期性のモデル”, 音響学会講演論文集, vol.1, pp.25–26, Oct. 2001.

[3] 酒向慎司，近藤重一，益子貴史，徳田恵一，小林隆夫，北村正，“唇動画像と音声によるマルチモーダルデータベースの構築”，音響学会講演論文集, vol.1, pp.221–223, Mar. 2001.