平成17年度 情報工学コース卒業研究報告要旨

末永 研究室	氏　　名	東 郷 高 浩
卒 業 研 究 題 目	バイモーダル音声認識における 画像情報利用に関する基礎的検討	

近年、実環境下での音声認識インターフェイスの重要性が高まっている。しかし、車内や人混みなど雑音の多いところでは、音声情報のみを用いた音声認識では認識率が低下し、雑音の程度や用途によっては実用的でない。そのため、様々な実環境の雑音下における音声認識率を向上させるために、映像情報を組み合わせたバイモーダルな音声認識手法が盛んに研究されている。映像情報は音声雑音の影響を受けない情報であり、音声認識において有効な情報源になり得ると考えられる。

映像情報の利用法として、音声認識に対して補助的に使用する発話区間推定等の手法と、映像から取得した特徴量を発話認識に使用する音声認識手法が挙げられる。前者は発話区間を正しく取得することで認識精度を上げるというものである。ただし、認識段階における雑音の問題は残る。後者では適切に映像特徴量と音声特徴量を組み合わせることにより、音声雑音による認識率低下を抑えることが出来る。しかし、映像特徴としてどのような特徴が有効なのか十分に調査されていない。

そこで本研究では、音声情報の特徴量と組み合わせる映像情報の特徴量として、1) 主成分得点，2) 正規化相関値, 3) 低輝度画素数について比較を行い、その有効性について検討する。1) はバイモーダル音声認識でよく用いられている特徴量で、口周辺領域の固有画像と入力画像から求める。2) 及び 3) についてはこれまでバイモーダル音声認識で利用されていない特徴量である。2) はフレーム間における口周辺領域の正規化相関値で、正規化されているため照明変動に強く、口が動いているかどうかの特徴を表す。3) は発話者の口腔内が周囲領域と比較して一般に暗いことを利用し口周辺領域におけるある閾値以下の画素数を特徴量としたもので、口の開き具合を表す。

認識率評価には、室内で撮影された数字発話タスクのバイモーダル音声認識用データベースである AURORA-2J-AV を使用した。音声情報の特徴量として MFCC を使用し、映像の入力には事前に切り出した口唇領域フレーム画像に映像雑音としてガンマ補正により照明変動を加えたものを使用する。各特徴量の一次及び二次差分値も同時に使用する。認識モデルには HMM を使用し、音声と映像の特徴量を分けたマルチストリーム HMM として音声と映像特徴量の重みを調整しながら認識結果を得た。実験の結果、正規化相関値を付加した場合は SNR10dB 以上において主成分得点を付加した場合に比べて認識率が平均約 4%高かった (表 1)。しかし、SNR5dB 以下の雑音が多い場合では正規化相関値より主成分得点を付加した場合の方が認識率が高かった。車内での雑音が SNR10dB 程度であることを考慮すると一般的な雑音下では有効であると考えられる。

表 1 映像特徴量付加による音声認識結果 (挿入ペナルティ調整後認識率%Acc)

SNR	映像特徴量				
	付加なし (音声のみ)	主成分得点	正規化相関	低輝度画素数	主成分得点 +正規化相関
20dB	87.85	91.06	94.00	88.85	90.60
15dB	84.95	89.24	94.03	78.76	89.13
10dB	75.66	84.83	88.71	67.72	84.70
5dB	54.05	72.67	71.90	53.10	73.14
0dB	17.13	57.79	51.66	29.68	59.24
-5dB	-9.36	48.08	22.98	12.44	50.97