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*

◎酒向 慎司　近藤 重一†　益子 貴史†　徳田 恵一　小林 隆夫†　北村 正　(名工大, †東工大)

1. はじめに

近年，聴覚情報だけでなく，顔画像などよる視覚情報を同時に取り扱ったマルチモーダルインタフェースに関する研究が盛んに行われている．

ところで，隠れマルコフモデルのような統計的手法に基づいた音声認識または音声合成に関する研究の成果は，大規模な音声データベースの整備に依るところが大きく，様々な研究用の大規模音声データベースが普及している．従って，視聴覚情報を用いた音声認識・合成の研究を推進するためには，マルチモーダル音声データベースの整備が不可欠であり，様々なデータベースが構築されつつある [1].

このような背景から，東京工業大学および名古屋工業大学の研究グループにおいても，音声と唇動画像から成るマルチモーダルデータベースの整備を進めようとしている．本文では，これを研究用データベースとして公開することを目標とし，その作成過程と仕様について述べる．

2. データベースの構築

2.1 収録方法

本データベース構築の概略を図1に示す．なお，作成に関する作業の内，撮影に関する部分は東工大小林研究室において行われ，その後のラベリングなどの作業は名工大北村研究室において行われた．データーベースの収録は，顔下部(鼻から顎まで)の顔画像を正面から家庭用ディジタルビデオカメラで撮影し，並行してDATデッキを使用してマイクから音声を録音した．ビデオカメラは，2.4で述べる理由により，90° 回転させて用いる．DVカメラにおいても音声のディジタル録音は行われているが，サンプリング周期，量子化精度の高いDATにより収録された音声データをデータベースとして使用する．収録時に使用した機材を表1に示す.

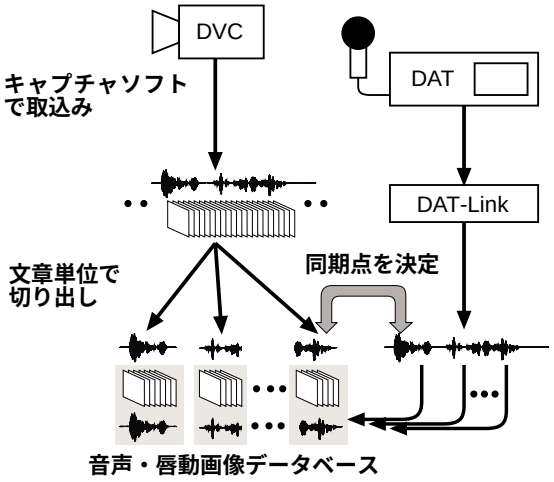

図 1．データベース構築の手順

表 1．収録に使用した機材

DVカメラ	SONY TRV-PC7
DATデッキ	SONY TCD-D7
マイク	SONY C-355

なお，データベース用のテキストは，ATR日本語データベースの音韻バランス文503文章 [2] で用いられているもの採用し，話者は放送研究会に所属する男子学生1名とした．その他，収録時の条件としては以下のようなものが挙げられる．

- 顔表面に影ができないように左右からスポットライトをあてる．
- カメラのフォーカスは自動とし，その他の補正等は行わない．
- DV，DATはともにSPモード (高品質) で記録する．
- 発話の前後には口を閉じるように指示.

2.2 データの取り込み

DVカメラによって収録されたNTSC DVフォーマットの動画像データは，PC用のビデオキャプチャーソフト (SONY DVgate Motion) を使用して，動画像トラックが720×480ピクセル (直方画素)，24bit RGBカラー画像，29.97フレーム/秒の非圧縮形式，音声トラックが16bit PCM(記録時は12bit)，サンプリング周波数32kHzのステレオ音声として，計算機に取り込んだ．さらに，ビデオ編集ソフト (Adobe Premiere) を使用して，文章単位による切り出しを行った．ただし，口の開き始めや閉じる時刻は，実際の発声の始まりや終りの時刻とは大きなずれがあるため，文頭・文末の口の開閉を基準として手作業で切り出しを行った．

一方，DATで収録した音声はDAT-Linkを用いて，標本化周波数48kHz，16bit PCM，モノラル音声として計算機に取り込んだ．この音声データとDVカメラのデータとは同期が取れていないため，両者の時間的な位置合わせを行う必要がある．本データベースでは，DVカメラの音声データと，DATの音声データを比較して同期点を調べ，動画像とDATの音声データとの同期を取った．同期点の決定は，標本化周波数を48kHzに再サンプリングしたDVカメラの音声データに対して，DATの音声データを1サンプルごとにずらしながら内積をとり，最大となる点と定め，文章ごとにこれを決定した．なお，再サンプリングの際には，時間的な遅延が生じないように配慮している．

得られた同期点を基準として，切り出された動画像に対応するDATの音声を切り出し，これらを音声・唇動画像データとした．動画像部は，非圧縮の

∗ Construction of a Japanese Multimodal Database.

By Shinji Sako, Kondo Shigekazu†, Takashi Masuko†, Keiichi Tokuda, Takao Kobayashi† and Tadashi Kitamura (Nagoya Institute of Technology, †Tokyo Institute of Technology)

24bit RGBカラー画像，720×480ピクセル(直方画素)，29.97フレーム/秒，音声部は，モノラル 16bit PCM，標本化周波数48kHzとなる．全503文章の合計は約50分，動画像データの容量は非圧縮の状態で約90Gバイト(全89726フレーム)となった．その他の付属データとして，次節で述べるラベルファイル，閲覧用のQuickTimeムービー(約340Mバイト)，フレーム毎の鼻孔中心位置座標などを作成した．

2.3 音素ラベリング

手作業によるラベルデータの作成は，専門的な知識を必要とし，大量の音声データを一定の品質でラベル付けを行うことは容易ではない．そこで，本データベースのラベルデータは，不特定話者のmonophone HMMを使い，Viterbiアルゴリズムによるセグメンテーションの結果に基づき音素境界を半自動的に作成した．不特定話者モデルは，収録した文章と同一のA-TRバランス音韻文503文章の男性話者6名分の音声データを用いて学習した3状態のleft-to-right monophone HMMとし，モデルの学習などはHTKを使用した．また，音響単位の分類はATR音声データベースに付属するラベルデータのうち，破裂音内の閉鎖区間などが考慮されたイベント層表記[3]のものを採用し，さらに閉鎖区間・入り渡り記号を環境依存としたより詳細な単位でモデル化を行っている．

なおセグメンテーションの際，ポーズを除いて出現する音素列は文章ごとに決まるが，音素列からイベント層記号列への変換は一意に決定できない場合があるため，ポーズの有無に加えて音素内のイベント層記号の並びには若干の自由度をもたせてある．

このようにして得られたイベント層レベルの記号列から自動変換によりATR音声データベースの音声表記層とイベント層のラベルに相当するラベルデータを作成し，データベースに添付することとした．

2.4 画像フレームのインタレース処理

キャプチャされた動画像データはインタレース走査されているため，1つの画像フレームは，水平方向の偶数番目と奇数番目のスキャンラインからなる2つのフィールドが含まれている．本来はこれらのフィールドは，フレーム周期の半分の周期で交互に書き換えられて再生されるものである．そのためキャプチャーされた動画像では，図2のように唇の開閉時などの比較的動きの激しいフレームにおいて，書き換えの処理が唇の動きに追い付かずに縞模様のようにずれた画像が生じてしまう．

そこで，インタレース走査されたフレームを利用して，スキャンラインの偶数列と奇数列を交互に分けて2枚の画像に分割することで，このような縞模様を解消し，さらに2倍の時間解像度をもった画像データを得ることが可能である．ただし，1ライン分の位置ずれが起ることに注意する．また，垂直方向の解像度が半分になるため，原画像と同じ縦横の画素数比にするためには，水平方向に対して2倍のダウンサンプリングもしくは垂直方向に2倍のアップサンプリングを行う必要がある．

また，通常のインタレース走査は画面の横方向に対して行われているが，唇の開閉の動きは縦方向の変化の方が大きい．インタレース走査により縦方向の空間解像度は実質的に半分になるため，その影響を抑えるためにカメラを90°回転させ，インタレース走査が顔の縦方向に行われるようにしている．

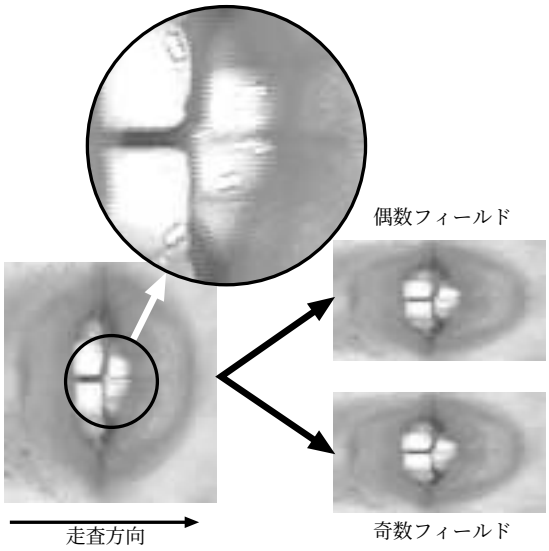

図 2．インタレース走査とフレームの分割

本データベースを使用して，音素のような詳細な単位で動画像をモデル化する場合には，継続時間の短い子音などでは，学習データとなる画像フレーム数の不足が問題となるが，時間解像度を高くすることにより，より多くの学習データ(画像フレーム)を得ることができる．これまでに，本データベースを利用した唇動画像HMMの学習[4]では，この手法を利用して59.94フレーム/秒の動画像データを作成し，唇動画像合成のためのモデルを音素単位で学習している．

3. むすび

ATR日本語バランス音韻文をテキストとして，男性話者1名による音声・動画像データベースを整備した．今後の課題として，韻律情報などの整備が挙げられる．なお，本データベースは，研究用として公開の予定である．

謝辞

本研究の一部は，科学研究費補助金奨励研究(A)(10780226)，萌芽的研究(11878064)，放送文化基金助成・援助金による．

参考文献

[1] Satoshi Nakamura, “Overview on Recent Activities in Multi-Modal Corpora,” COCOSDA Workshop, Oct. 2000.

[2] 阿部 匡伸, 匂坂 芳典, 梅田 哲夫, 桑原 尚夫, “研究用日本語音声データベース利用解説書(連続音声データ編),” ATR日本語音声データベース付属資料, 1990.

[3] 武田 一哉, 匂坂 芳典, 片桐 滋, 桑原 尚夫, “音声データベース構築のための視察に基づく音韻ラベリング,” ATR日本語音声データベース付属資料, 1988.

[4] Shinji Sako, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi and Tadashi Kitamura, “HMM-based Text-to-audio-visual Speech Synthesis – Image-based Approach,” ICSLP, vol.III, pp.25–28, Oct. 2000.