ＴＲ－ＳＬＴ－００７３

雑音環境下における合成発話動画像の評価

Evaluation of Lip-sync for Talking Head In Noisy Environments

前島 謙宣　　四倉 達夫
Akinobu MAEJIMA　　Tatsuo YOTSUKURA

森島 繁生　　中村 哲
Shigeo MORISHIMA　　Satoshi NAKAMURA

2004年3月31日

概要

著者等は，自然な発話顔アニメーションの合成手法を提案している．しかしながら，その性能に対する評価は課題として残されていた．発話顔アニメーションの性能は，(1) 読唇をできる程度に再現されているか，(2) 視覚的に自然であるか，(3) 音声と正確に同期しているかの3点により決定される．本稿では，まず雑音環境下において発話顔アニメーションと音声とを被験者に提示し，発話内容の聞き取り実験により (1) を検証する．次に (2) について，発話顔アニメーションの視覚的な自然さおよび，発話口形の滑らかさを5段階評価する．最後に (3) について，一定間隔で音声と発話顔アニメーションとの同期をずらしたものを被験者に提示し，同期のずれの主観値を調査するとともに，違和感の程度を5段階評価により評価する．加えて音声と発話顔アニメーションとの同期のずれが音声の知覚に及ぼす影響についても評価する．以上から，合成発話顔アニメーションの品質を評価するとともに，音声との自然な同期について検証した．

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目2番地2　TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai “Keihanna Science City” 619-0288,Japan
Telephone:+81-774-95-1301
Fax　　:+81-774-95-1308

1．はじめに

1.1　研究背景

近年，人間と機械との自然で豊かなインタラクションの実現を目的とした，ヒューマンコンピュータインタフェースの研究が盛んに行われている．人間のような見た目と振る舞いをする擬人化エージェントの実現は重要な課題の一つと言える[1]．エージェントには，その性質上ユーザである人間とFace-to-Faceの自然なコミュニケーションを行えることが強く望まれている．

ところで，Face-to-Faceの対話において，音声のような聴覚的・言語的な情報だけでなく，口の動きや表情といった視覚的・非言語情報も重要な役割を果たしている．一般に人間は，発話内容を知覚する場合，聴覚的な情報だけでなく，視覚的な情報，会話の話題，文章の前後関係，過去の経験といった情報を統合し，発話内容を認識していると考えられている．特に音声や口の動きには互いに強い相関があると考えられている．このような二つのモダリティ間の同期が取られていない場合，人は不自然な感覚を覚えたり，誤った知覚をする[2]．従って，エージェントには，自然かつ音声との同期が取れた正確な口の動きが要求される．自然なエージェント実現のための，合成発話顔アニメーションに関する研究は，既にいくつか報告がなされている[3～8]．

著者等は，従来より標準的な3次元顔モデルを個人の顔に適用するモデルベースの発話顔アニメーションの合成手法を提案しており，これを応用したビデオ翻訳システムを構築している[9][14]．ビデオ翻訳システムとは，音声翻訳だけでなく発話者の口領域の画像も翻訳するものであり，これによりマルチモーダルな翻訳を可能にしている．提案手法の利点は，顔が3次元モデルで表現されているため，移動・回転に制約がなく，発話口形を生成するためのルールを構築することが容易でかつ，特定の個人に限らず適用できることが挙げられる．

このような，発話顔アニメーションの評価には一般的に客観評価法[7]と主観評価法[12]が用いられるが，合成手法の性質上，客観評価を行なうのは困難であり，従来から主観評価により性能の評価を行ってきた．しかしながら，発話顔アニメーションの再現性，外観の自然性，音声との同期に関する厳密かつ定量的な評価は行なわれていなかったと言える．

1.2　研究目的

本研究では発話顔アニメーションの品質を以下，の3点について評価実験を行うことで検証した．

(1) 読唇をできる程度に再現されているか　（以後，発話顔アニメーションの再現性と呼ぶ）
(2) 視覚的に自然であるか　（以後，発話顔アニメーションの自然性と呼ぶ）
(3) 音声と正確に同期しているか　（以後，発話顔アニメーションの同期性と呼ぶ）

本研究では，まず雑音環境下において発話顔アニメーションと音声とを被験者に提示し，発話内容の聞き取り実験を行うことにより(1)を検証する．次に(2)について，発話顔アニメーションの視覚的な自然さおよび，発話口形の滑らかさを5段階評価する．最後に(3)について，ある一定間隔で音声と発話顔アニメーションとの同期をずらしたものを被験者に提示し，同期のずれの主観値を調査するとともに，違和感の程度を5段階評価により評価する．加えて音声と発話顔アニメーションとの同期のずれが音声の知覚に及ぼす影響についても評価する．これらの評価実験を通じて，著者等が提案する合成発話顔アニメーションの品質を評価すると共に，合成発話顔アニメーションと音声との自然な同期について検証した．

2．合成発話顔アニメーションの生成

本研究では，[9]に基づく手法により，合成発話顔アニメーションを作成した．手順を図1に示す．まず後述するデータベース中の各画像シーケンスに対して，その初期画像フレームを用いて，個人の3次元顔モデルを作成する．次に，映像中の人物の顔の動きを推定するために，個人用顔モデルより生成される3次元顔テンプレートを用いて自動顔トラッキング[9] [11]を行う．

さらに，発話内容を既知として音声を音素セグメンテーションすることにより，音素継続長情報を得る．ここで発話内容の音素表記から，それに対応する発話口形を基本口形データベースから取得する．また一方で音素継続長情報を用いて基本口形間の口形状の補間を行う．これはキーフレームアニメーションの原理に基づく．最終的に得られた発話している口領域モデルを，顔トラッキングにより推定された位置・角度へと埋め込むことにより，合成発話顔アニメーションが生成される．

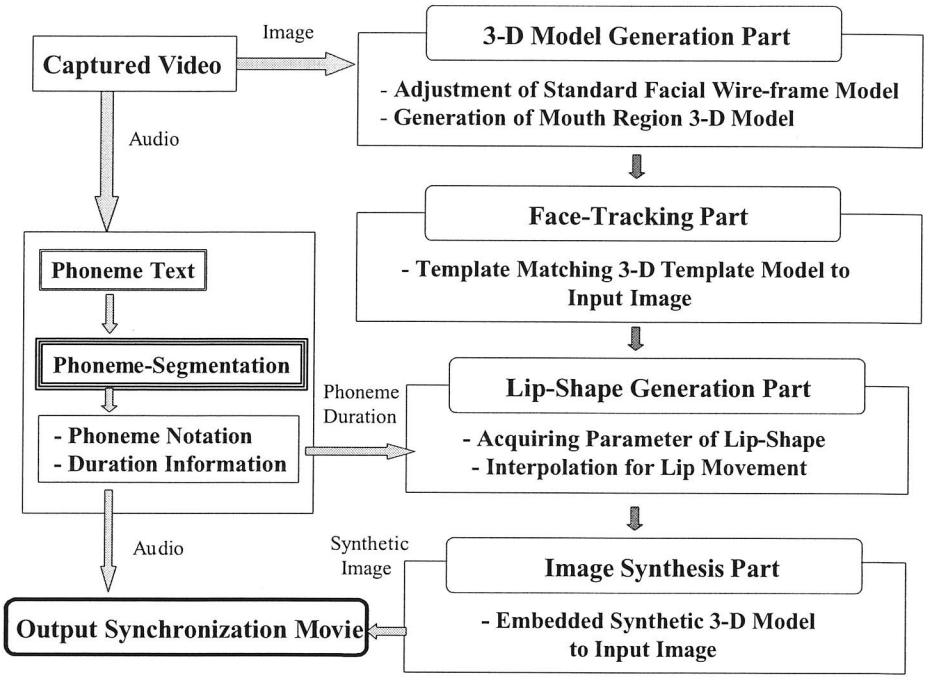

図1. 合成発話顔アニメーションの生成手順

2.1　3次元顔モデルの生成

人間の顔は，基本的な形状や構造は同じといってよいが，目，鼻，口等の各部位の形状や位置は個人に依存する．このためCGにより自然な顔を合成するには，対象となる人物の顔により忠実かつ演算量の少ない3次元顔モデルを構築する必要がある．そこで本研究では，標準ワイヤーフレームモデルを個人の顔へと整合することにより，個人の3次元顔モデルを生成した．

図2(a)は画像シーケンスの任意の1フレームから取得した原画像を示している．図2(b)は，原画像に対して標準顔ワイヤーフレームの整合を行った結果を示している．この整合には専用のGUIを用いて約1分程の作業を必要とする[10]．そして最後に，整合により得られた画像を3次元顔モデルにテクスチャマッピングを施すことにより，個人の3次元顔モデルが得られる．

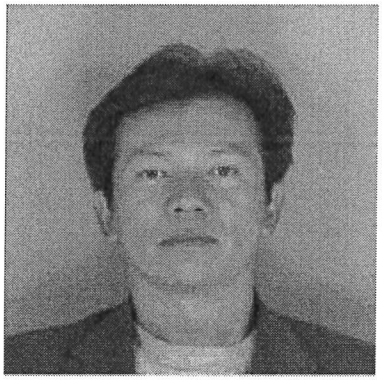

図2.a 現画像

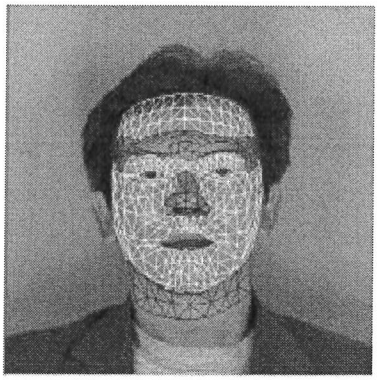

図2.b ビデオフレームへの整合

図2.c 取得した口領域モデル

2.2 自動顔トラッキング

合成発話顔アニメーションは,2.1で述べた3次元口領域モデルを画像上に埋め込むことにより生成される.しかしながら,3次元口領域モデルを,画像上のどの位置にどのくらい回転させて埋め込めばよいのかは明らかになっていない.このため,あらかじめ映像中の人物の顔の位置・角度を知る必要がある.本研究では,3次元顔テンプレートモデルを用いた自動顔トラッキングにより,映像中の人物の顔の位置・角度を推定した[11] [9].

2.3 音素セグメンテーション

本研究では,音声の発話内容は既知であるとして,音素セグメンテーションにHTK[13]を用いた.音響特徴は,,16kHzサンプリング周波数,フレーム長25mec,フレーム周期10msecで抽出された12次元MFCC,12次元⊿MFCC,⊿対数パワーを使用した.音響モデルには,話者非依存のIPA準拠のモデルを用いた.また,音素セグメンテーションに誤りがある場合は,音声と口の動きとの同期がとれなくなるため,専用のGUIツールを用いて手動で音素境界の修正を行った.様子を図3に示す

2.4 発話顔アニメーションの生成

2.4.1 発話口形の生成

人間が会話する際の動作の大きな部位として,唇,顎などが挙げられる.とりわけ,唇の動きは音韻と密接な関係を持つため,正確な制御が必要とされる.倉立らは,被験者の顔にマーカを置き,運動情報を計測している[5].このようなアプローチは,運動を正確に計測でき,柔軟な制御を行えるという利点がある.

本研究では,ビデオ翻訳システム[9]で使用されている発音記号をVISEMEに基づいて日本語について12種類に分類し,さらに無音状態を加えた13種類の基本口形を,基本口形データベースとして使用した(表1).基本口形は,ワイヤーフレームモデルの口領域のベクトル移動量を定義した口形パラメータと,3次元計測器を用いて計測することで得られる無音および発話時における口形状の3次元位置情報のベクトル移動量を用いて作成される.ここで,後者のベクトル移動量については話者数人分について計測し,その平均値を用いることにより個人性を除去している.また,本来VISEMEは2重母音[au] [ei]等に現れる口唇運動の情報まで定義されるのであるが,そのようなVISEMEは,二つのVISEMEから構成されるものとし,二つの基本口形を割り当てることにより表現した.前半に現れる口形については音素継続長の30%を,後半の口形については残りの音素継続長を経験的に割り当てた.これにより,話者に依存しない小規模なデータベースを構築している.例として図3に無音時と母音a発話時の口形を示す.

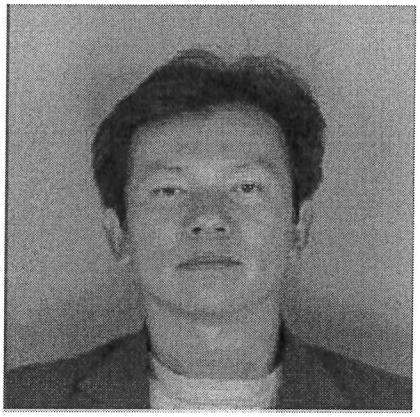

図3. a 現画像

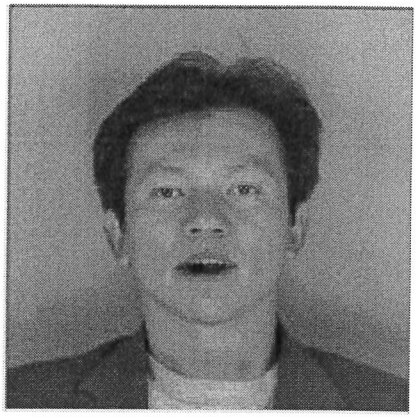

図3. b 母音a発話時の口形

表1. 音素とVISEMEの対応

VISEME No.	音素表記
1	/a/
2	/i/, /y/
3	/u/
4	/e/
5	/o/
6	/r/, /ry/
7	/b/, /p/, /m/, /by/, /py/, /my/
8	/t/
9	/d/, /n/, /ny/
10	/g/, /k/, /N/, /hy/, /gy/, /ky/
11	/f/
12	/j/, /s/, /z/, /ch/, /dy/, /sh/, /ts/
0	/#/ 無音

2.4.2 発話口形状の補間

基本口形データベースより取得された基本口形をキーフレームとして，キーフレームアニメーションを行う．ビデオレートに依存したキーフレーム以外のビデオフレームにおける発話口形状は，基本口形データベースには存在しておらず，このままでは音声と発話口形状との完全な同期を得ることができない．そこで，音素継続長情報を利用して基本口形間の発話口形状の補間を行う．

音素の発声開始時は必ず基本口形状を構成しているとして，これをキーフレームとした．そして音素継続時間の始点では，基本口形状を構成する格子点のベクトル移動量の重みを100%とし終点では0%になるように，また後続する音素についてもベクトル移動量を始点から終点へ0%から100%となるようにし，二つのベクトル移動量の加算を行うことにより発話口形状の補間を行った．

本研究では，口形状の補間に線形補間法，正弦波補間法を用いた．両者の補間法の概念図を図4a, 4bに示す．一般的にこのような補間を行う際には，トライフォンを考慮することで再現性が増すと考えられている[4]．しかしながらここに示すような補間法により，3つの基本口形のベクトル移動量を加算することでトライフォンに対する口形を表現するのは困難であり，演算量も増大するという問題もある．そこで本研究では，後続する音韻のみを補間の対象とした．

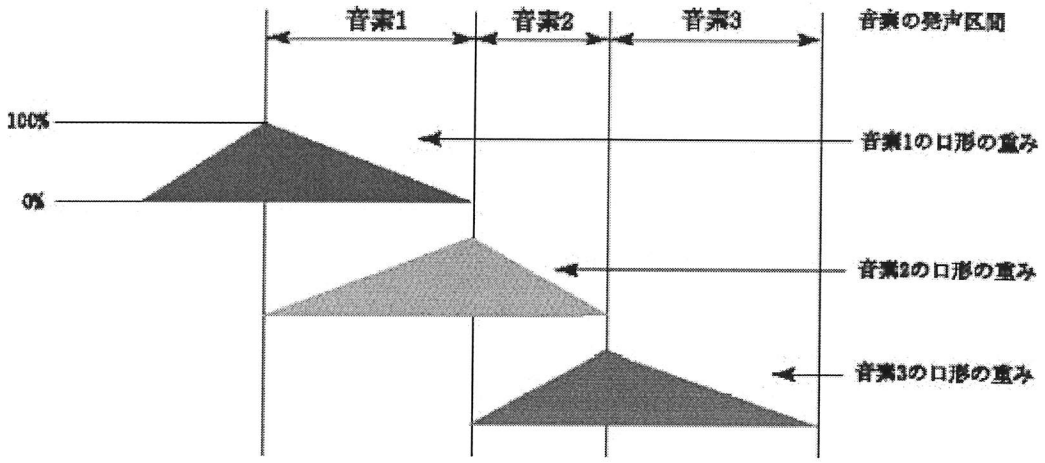

図4.a　基本口形の線形補間によるベクトル移動量の重み付け

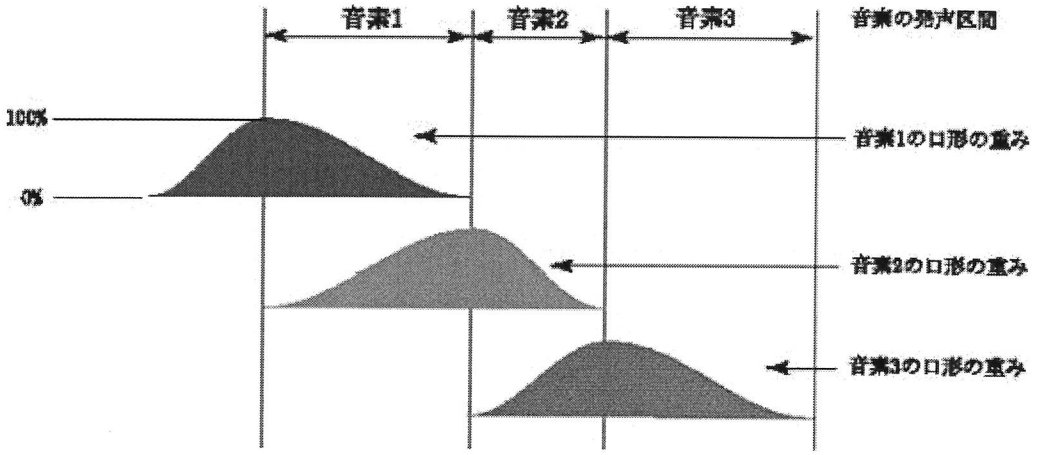

図4.b　基本口形の正弦波曲線補間によるベクトル移動量の重み付け

2.4.3　合成発話顔アニメーションの生成

2.4.2で得られた3次元口領域モデルをビデオフレーム画像に重ね合わせることで合成発話顔アニメーションが生成される．しかしながら，単に口領域モデルを画像に重ね合わせただけでは，モデルと画像との間に境界が発生してしまう．これを避けるために，モデルの境界に対してαブレンディングを施すことにより境界を除去する．こうして得られた口領域モデルをビデオフレーム画像へ重ね合わせることで境界の無い自然なアニメーションを作成することができる．

3. 評価対象の作成方法

本章では，実験に用いる評価対象の作成方法について述べる．評価対象に用いる発話顔アニメーションには，10進4桁の数字列を使用し，雑音環境下において発話顔アニメーションの存在が音声の明瞭度改善にどの程度寄与しているのかを調べるため，音声には，-6, -20, -30dBのホワイトノイズを付加した．評価実験の詳細についてはここでは述べず，次章以降で説明する．

3.1 評価対象動画像の撮影及び編集

評価対象用動画像の作成手順は以下のとおりである(図5)．

1. 発話の開始を知らせる合図(以後，発話トリガと呼ぶ)となる「番号は」と，0から9までの発話をDVカメラで撮影する．
2. 撮影された動画像から各単語(数字)が発話されている区間の画像シーケンスの切り出しを行い，これをデータベースとする．
3. 各単語の画像シーケンスに対して合成発話の動画像を作成し，データベースに保存しておく．
4. 評価対象の発話内容となる4桁の数字列を，乱数を用いて生成する．
5. データベースから発話トリガと，発話内容に該当する数字の動画像と音声を取得し，それらを接続する．
6. 5で生成された動画像に対して，音声に-6, -20, -30dBのホワイトノイズを付加する．

以上により評価対象となる発話顔アニメーションが作成される．ここで，自然発話顔アニメーションに関しては，自然発話の画像シーケンスを，音声のみの場合は，背景を全て黒にした画像を用いている．音声のみ場合，背景が黒であることを除いて，自然発話顔アニメーションと同様の手法により作成される．

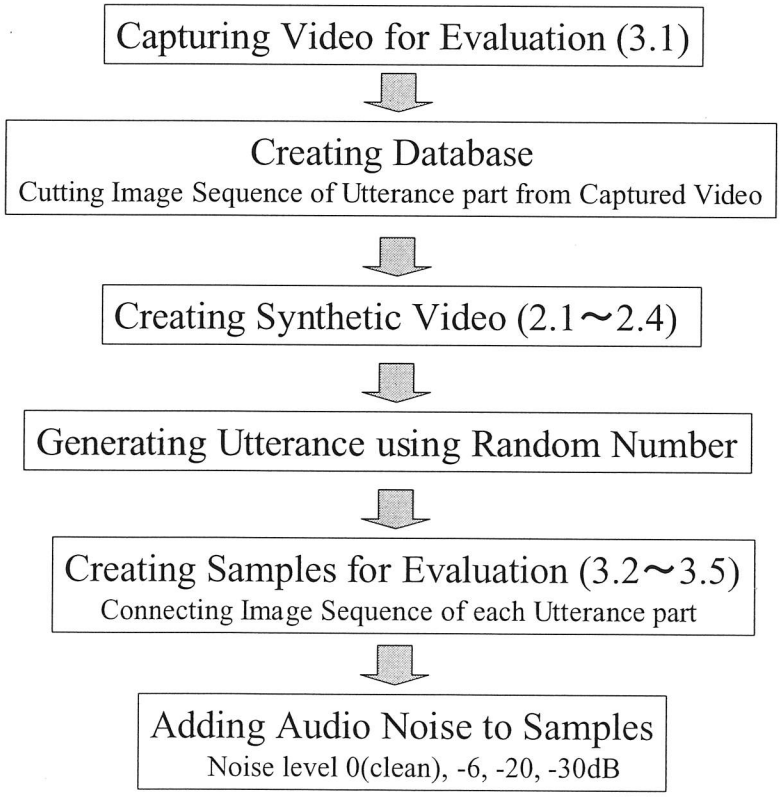

図5. 評価対象動画像の作成手順

3.2　自然動画像の撮影

まず，自然動画像として，正面から被験者が発話している様子をDVCAM(SONY DSR-PD-150)を用いて撮影した．このときフレームレート29.97[fps]，サンプリングレート48[kHz]である．発話内容は，「番号は」と0から9までの数字の離散発話である．また，カメラー被写体間の距離は3mとし，照明は被写体の顔に影ができないよう一定とした．このときの撮影環境を図6に示す．ここで，撮影された動画像に対して，あらかじめ単語（数字）単位に切り出しを行っておく．これをデータベースとして，以降で説明する評価対象の作成を行った(図7)．

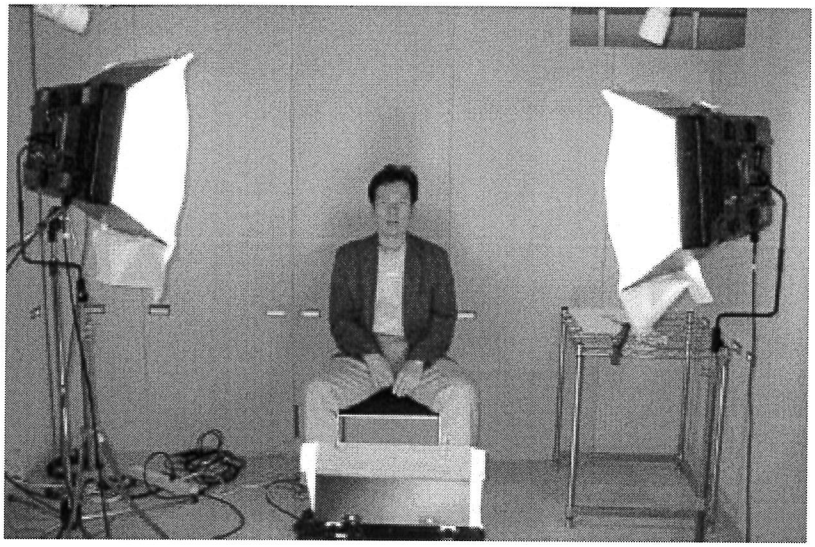

図6. 撮影風景

3.3　自然発話顔アニメーションの作成

評価対象として用いる自然発話顔アニメーションは，発話内容となる4桁の数字列を，乱数により生成後，データベース中の画像シーケンスと音声とを，数字列に従って各単語(数字)単位で接続することにより作成した(図7)．例えば，図7において，数字列が「0941」となった場合，まず初めに，発話内容の開始を知らせる発話トリガとなる「番号は」，次に，「0」「9」「4」「1」と順に画像シーケンスと音声が接続されて行く．ここで，接続された映像フレームおよび音声の境界付近における不連続性が問題となる．そこで，映像フレーム間の境界における映像の不連続性を軽減するため，撮影時に被験者に極力顔以外の部分を動かさないよう，そして単語発話後は必ず口を閉じるように依頼した．さらに，境界間の二つの映像フレームに対してαブレンディングにより映像フレームを混合し，生成された画像を新たなフレーム画像とした．本研究では，画像境界の前後2フレームをαブレンディングの対象とした．

また，音声の境界については聞き取り実験の際に雑音が付加されるため，特別な処理は行わず，数字列に従い単に音声波形を接続した．

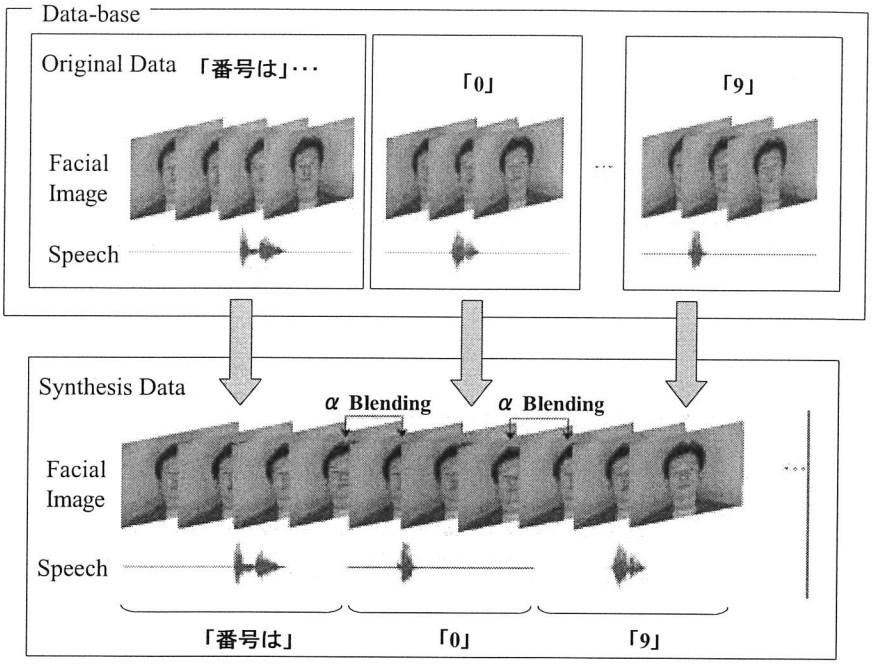

図7. 評価対象の作成

3.4 合成発話顔アニメーションの作成

評価実験の際に用いる合成発話顔アニメーションは，データベース中の各画像シーケンスに対して2章の手法により作成され，データベース中に保存される．そして乱数により数字列が生成された際に，3.2節と同様の手法で画像シーケンスと音声が接続されることで合成発話顔アニメーションが作成される．また実験の際に，被験者に合成であることに気づきにくくさせるため，発話トリガとなる「番号は」に関しては，自然発話の画像シーケンスを用いた．

3.5 不一致発話顔アニメーションの作成

音声と発話顔アニメーションの同期の重要性を検証するために，音声と発話口形との同期が，一定時間ずらされた発話顔アニメーションの作成を行った．本論文では，今後この発話顔アニメーションを不一致発話顔アニメーションと呼ぶことにする(本論文では，時間的な同期の不一致にのみ言及することとする)．不一致発話顔アニメーションは，自然・合成発話顔アニメーションにおける音声と発話口形との同期を，Adobe Premiereを用いて±33msec間隔で±500msecまでずらすことにより作成した．

4．実験1. 合成発話顔アニメーションの再現性の評価

合成発話顔アニメーションの発話口形が実際の発話口形に対して，正確に再現できているかどうか評価実験を行なった．実験の被験者は，大学生20代男性13名であった．

4.1 実験方法

自然音声(0dB)および，-6dB，-20dB，-30dBのホワイトノイズを付加した音声と，自然発話，および2種類の口形状補間法を用いて作成した合成発話顔アニメーションと，全体が黒の画像(音声のみの状況を意図的に作り出すため)を被験者に提示し，雑音に埋もれた音声の発話内容の聞き取り実験を行った．実験の被験者は20代男性13名である．

また，より多くの人に評価を依頼すべく，CGIにより評価実験専用のWEBサイトを開設し，実験環境を満たす環境下にいれば時間と場所を問わずに評価実験を行えるようにした．本WEBサイトの様子を図8に示す．

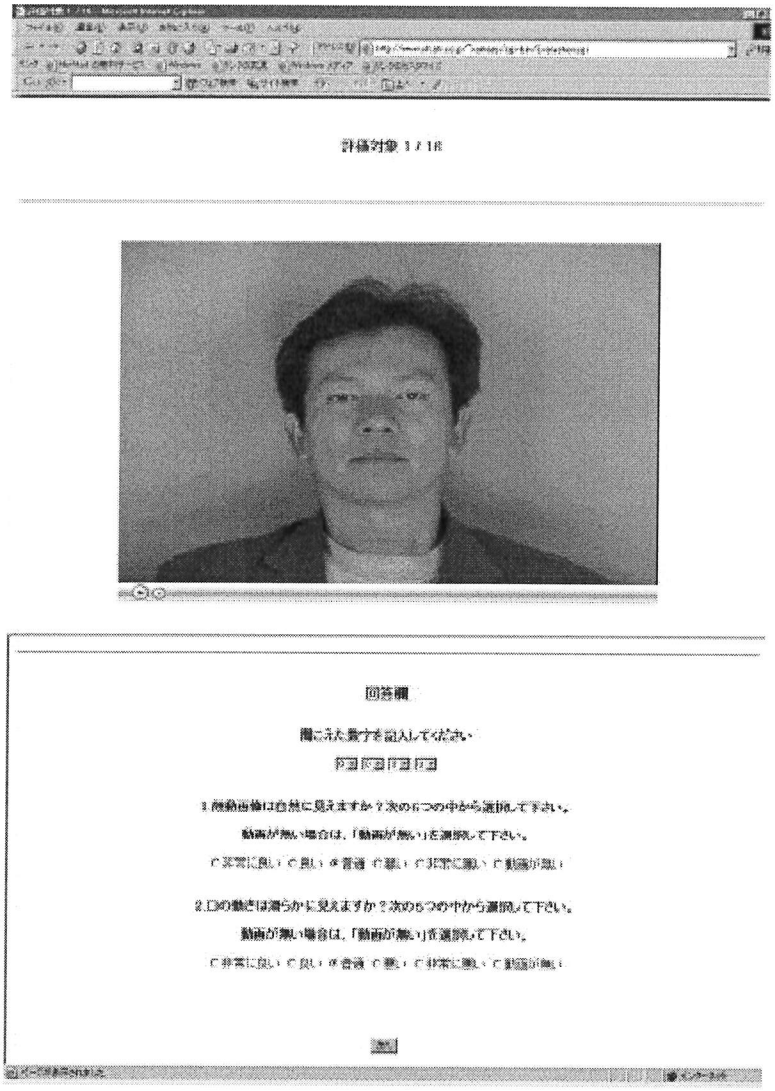

図 8. 評価実験 WEB サイトの様子

被験者間で実験環境に際の出ないよう，評価対象については乱数を用いてランダムに提示するようにした．また，音量に関してはCGIにより制御し，さらに実験を行う上で必要な実験環境や実験条件を提示し，遵守させることによりこの問題を解決した．以下に本実験に必要な実験環境と，実験条件を示す．被験者は，提示される評価対象を視聴し，発話内容である4桁の数字列をフォーム上に記入する．

実験環境

- OS Windows98 以上
- CPU PentiumⅢ 1GHz 以上
- Memory 256MB 以上
- Windows Meida Player6.0 以上をインストール済み
- ADSL以上の高速回線に接続されていること

実験条件

- 実験の際は必ずヘッドホンを使用すること
- マシン音量を必ず最大にすること
- 評価対象は必ず1回のみ視聴すること

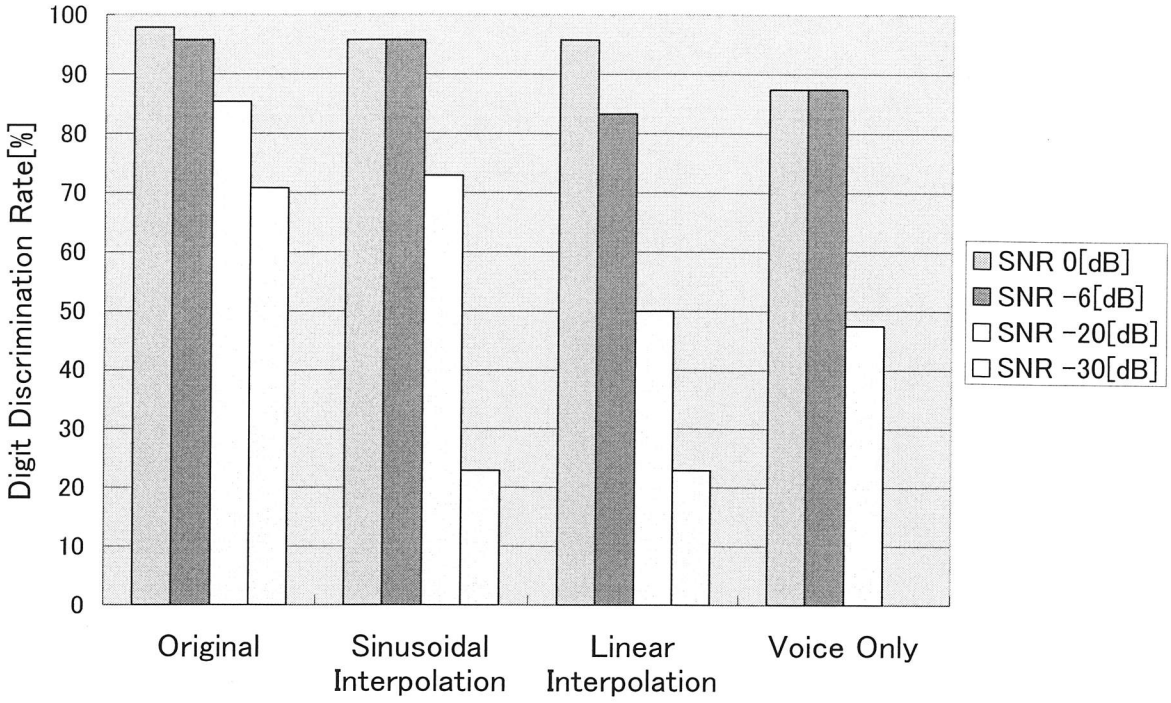

図 9. 合成発話顔アニメーションの再現性の評価

4.2 考察

実験結果を図 9 に示す．ここで，被験者が発話内容である 4 桁の数字列のうち幾つ聞き取ることができたかを表す指標として数字識別率を定義する．図 9 より自然・合成発話アニメーションの両方について，音声のみを被験者に提示した場合よりも，数字識別率は約 10% 程度高く，発話アニメーションを提示することにより，音声の明瞭度が改善されていることが伺える．

特に雑音レベルが -30dB の場合，被験者にはほとんど雑音しか聞こえていない状態であるが，発話アニメーションが存在することにより，自然発話の場合では 70% 近い改善がなされている．これは，発話内容が 10 進 4 桁の数字に限定されていて，かつ提示されている発話アニメーションから読唇により発話内容を推定し，認識した結果であると言える．つまり，提示されている発話アニメーションの再現性が高ければ，読唇により雑音に埋もれた音声の明瞭度が改善されると考えられる．

したがって合成発話顔アニメーションの数字識別率が，自然発話に近いならば，合成発話顔アニメーションの口形の再現性も自然な状態に近いということが言えるであろう．ここで，正弦波補間，線形補間 [9] により作成された 2 種類の合成発話顔アニメーションに着目すると，-30dB において，線形補間法と正弦波補間法は，22.5% となっている他は，正弦波補間法が全体的に数字識別率が高いものとなった．しかしながら，自然発話の場合と比較すると未だに 50% 程の開きがあり，口形の再現の改善が必要であると考えられる．

5．実験2. 合成発話口形の自然性の評価

4章の実験では，発話顔アニメーションの再現性のみについて評価を行なった．ここでは，合成発話顔アニメーションにおける発話口形の自然性について評価する．

5.1　実験方法

4章での実験の際に，自然・合成発話顔アニメーションに対して，視覚的な自然さ，口の動きの滑らかさについて「非常に良い，良い，普通，悪い，非常に悪い」の5段階評価を行った．評価実験の被験者は，大学生20代男性13名である．実験には4章で説明したWEBサイトを用い，評価結果は被験者の主観値の平均であるMOS(Mean Opinion Score)を用いた．

5.2　考察

発話顔アニメーションの自然性について考察する．図10に発話顔アニメーションの視覚的な自然さ，図11に発話口形の滑らかさに対する評価結果を示す．図より明らかに合成発話顔アニメーションは，自然発話顔アニメーションと比較して，視覚的な自然さおよび発話口形の滑らかさの点で劣っている．原因には，口の動きが自然発話と比較して不自然であることと，口内モデルや舌モデル，歯モデルの視覚的な不自然さなどが挙げられる．特に後者は，顔モデルに張られているテクスチャが実際の人間のものであるため，顔モデルと口内モデルとの視覚的な不整合が不自然さを強調していると考えられる．

また，2種類のキーフレーム補間法である，正弦波補間法と線形補間法を比較したところ，正弦波補間法がやや自然であるとの回答が得られた．これは線形補間法が実際の人間の口の動きと比較するとやや機械的であるのに対し，正弦波補間法が人間の口の動きに近い表現が行えているためであると考えられる．

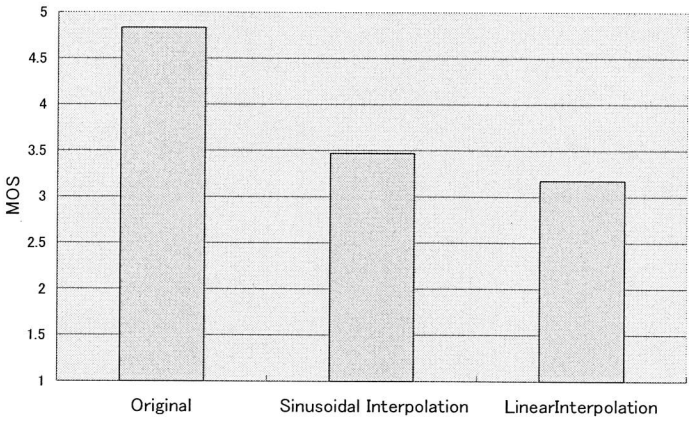

図10. 発話顔アニメーションの視覚的な自然さ

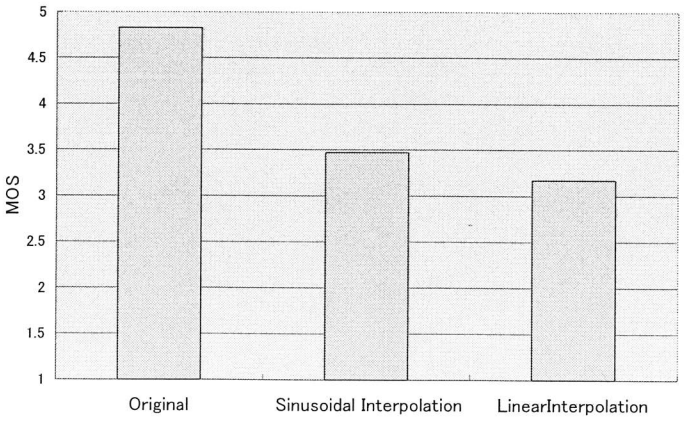

図11. 発話口形の滑らかさ

6. 実験 3. 音声と発話顔アニメーションの同期に対する評価

音声と発話口形との同期が正確かつ自然に取られているのか検証する．同期には，音韻的な同期（音素とそれに対応する発話口形の一致）と時間的な同期（音素と口形の時間的なタイミングの同期）とが存在するが，実験で用いる発話内容の音韻情報が既知であるため，ここでは時間的な同期についてのみ議論する．

6.1　音声と発話顔アニメーションとの主観的な同期のずれ

6.1.1　実験方法

自然音声と，±30msec 間隔で±500msec まで同期がずらされた自然・合成の不一致発話顔アニメーションを被験者に提示し，被験者にどの程度同期がずれていたか主観値で回答させた．ここでは，音声よりも口の動きの早い場合を＋，遅い場合を－とした．実験の被験者は，大学生 20 代男性 8 名および社会人 20 代男性 1 名である．

主観値は，自然な同期状態を 0，音声よりも口の動きが早く，同期のずれが +500msec の場合を +100，逆に音声よりも口の動きが遅く，同期のずれが -500msec の場合を -100 とし，被験者に自然な同期からのずれを主観値により回答させた．さらに実験の最後にどの程度同期がずれると違和感を感じたかそのずれの程度を同様に回答させた．図 12 ～ 16 に実験に使用した実験シートを示す．

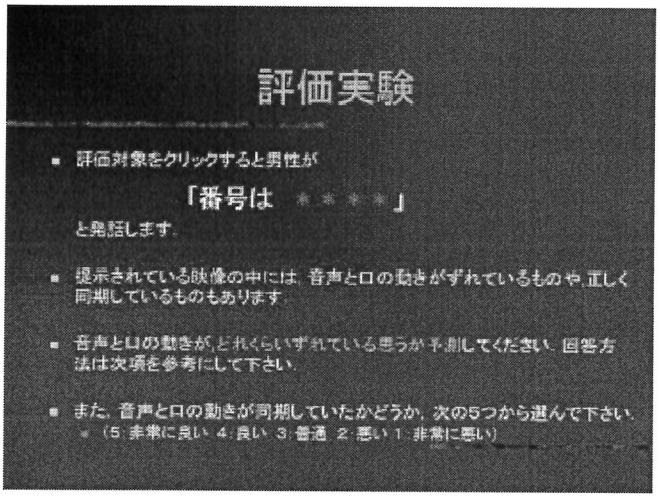

図 12　評価実験の実験内容説明

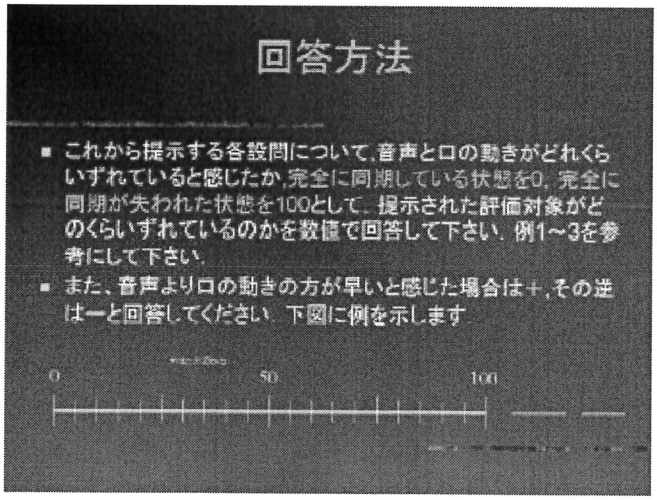

図 13　回答方法の説明

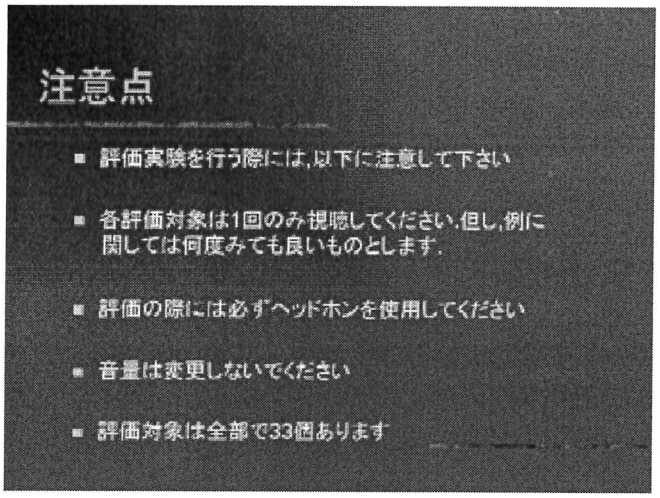

図 14　実験を行う上での注意事項

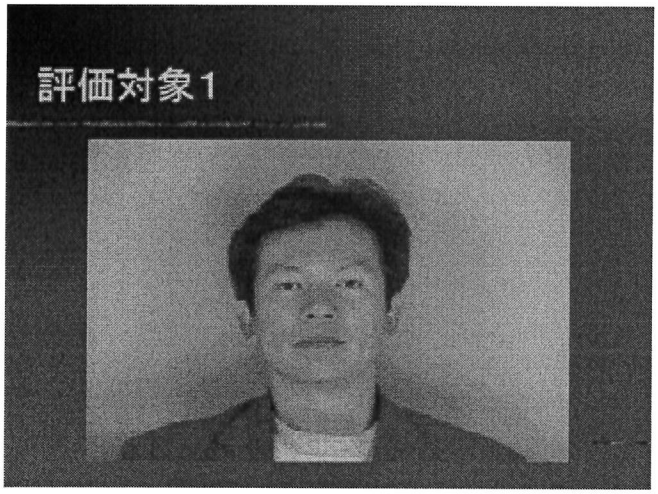

図 15　評価対象例

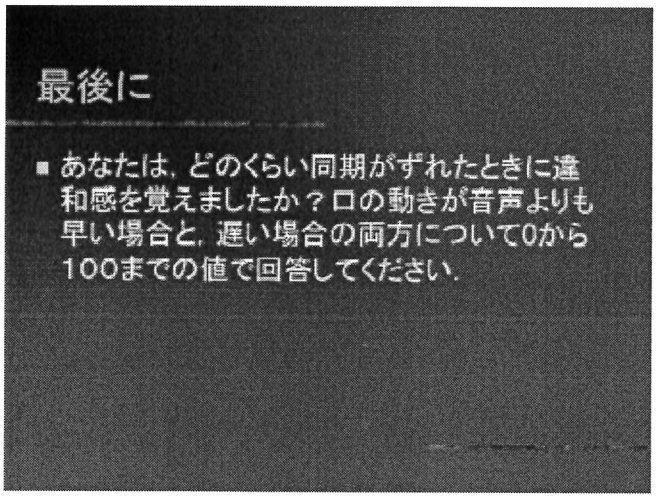

図 16　違和感の調査

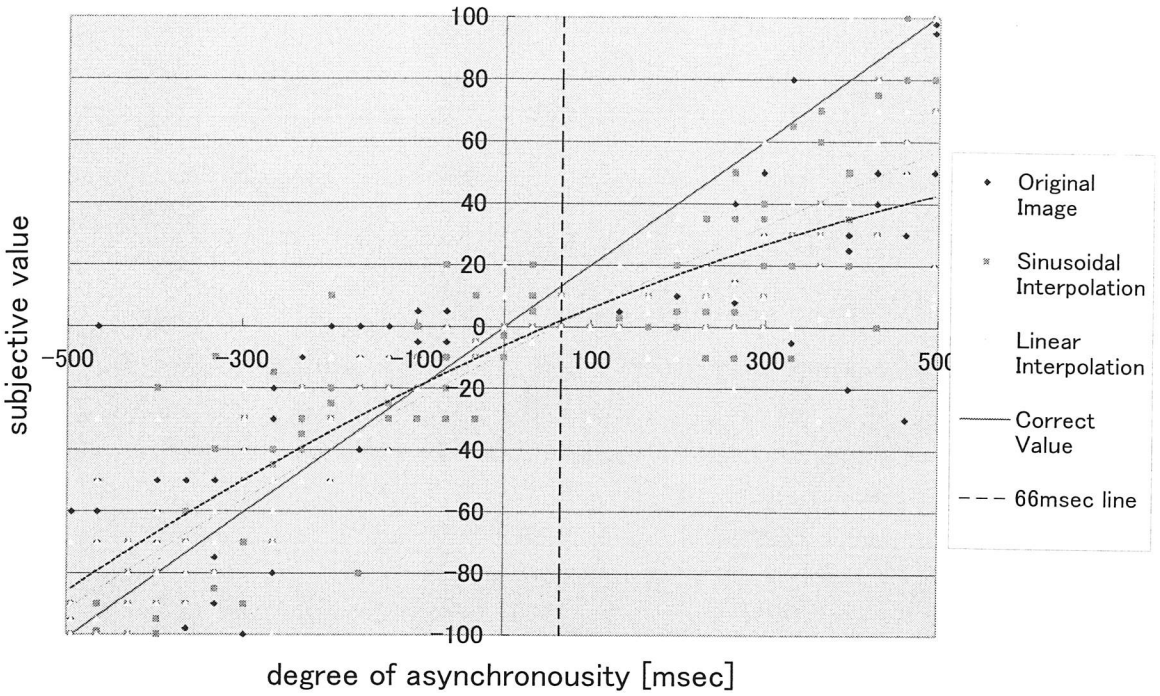

図17 音声と発話顔アニメーションの主観的な同期のずれ

6.1.2 考察

図17に，音声と発話顔アニメーションの主観的な同期のずれに対する結果を示す．各点は，発話顔アニメーションにおける被験者の回答の分布を，曲線は，分布に対する2次の最小自乗曲線を表している．

図17に注目すると，音声よりも口の動きの方が早い場合においては，被験者にはさほどの違和感がなく，実際の同期のずれよりも小さいと感じている．これに対して，音声が口の動きよりも早い場合は，実際の同期のずれとほぼ同じ値となっている．これは人間が発話するときには，調音器官が開口等の動作をした後に音声が発せられるという事実に起因するものと考えられる．実際，+66msec付近に同期タイミングがあると被験者は感じており，これを裏付けている．

6.2　音声と発話顔アニメーションの非同期による違和感

音声と発話顔アニメーションとの同期がずれた場合に，人間の知覚に与える影響について検証した．実験の被験者は，大学生 20 代男性 17 名および社会人男性 20 代男性 2 名である．

6.2.1　実験方法

実験には前節と同様に，自然音声と自然・合成の不一致発話顔アニメーションを用い，提示された評価対象が自然であるかどうか「非常に良い，良い，普通，悪い，非常に悪い」の 5 段階評価を行った．実験には，図 12 ～ 16 のシートを用いた．

6.2.2　考察

図 18 における点は，被験者の発話顔アニメーションの自然性に対する評価結果の分布を表しており，曲線は，分布に対する最小自乗曲線を表している．

図 18 に着目すると，被験者間で同期のずれに対して感じる違和感の程度にはばらつきがあり，図 17 のように，発話顔アニメーションと音声との同期が 1 番自然に取られていると感じる明確な同期位置を決定することが困難な結果となった．しかしながら分布の傾向は，口の動きが早い場合に傾倒しており，今後より多くの被験者に対して評価実験を行うことで，図中の最小自乗曲線のような分布になるのではないかと考察する．

次節では，図 18 における最小自乗曲線を目安として，被験者が「非常に良い，普通，非常に悪い」と回答している 5 つの同期タイミングについて着目し，同期のずれが音声認識におよぼす影響について検証する．

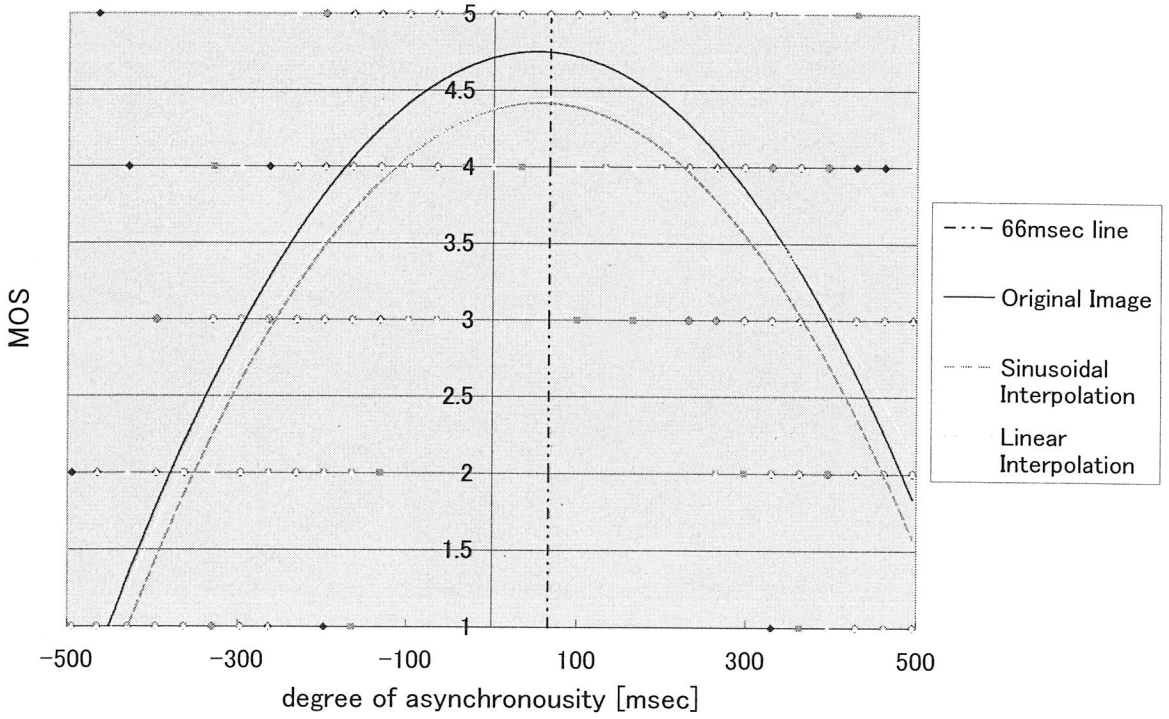

図 18. 音声と発話顔アニメーションの同期のずれによる違和感

6.3　同期のずれが音声認識に及ぼす影響

音声と発話顔アニメーションとの同期がずれたときに，人間の音声認識におよぼす影響について検証した．

6.3.1　実験方法

前節の結果から，被験者が「非常に良い，普通，非常に悪い」と回答している5つの同期タイミング(-462, -297, +66, +396, +495msec)について，4章と同様の実験を実施した．但し，雑音レベルは読唇による効果が生かされ，かつ音声を聞き取れる雑音レベルである-20dBに限定して実験を行った．図19～21に示す．

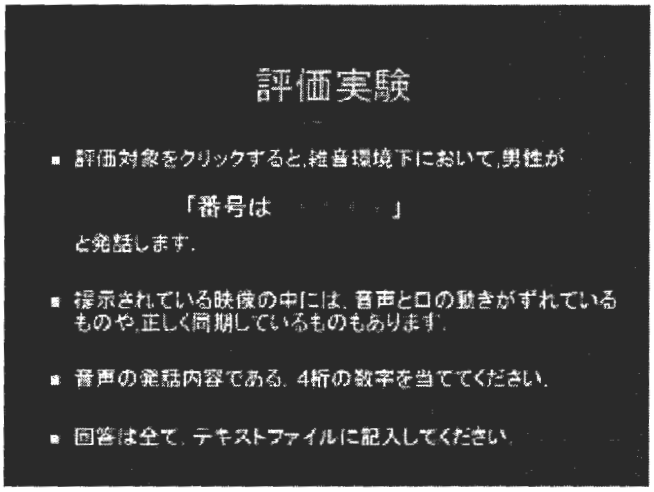

図19　評価実験の実験内容説明

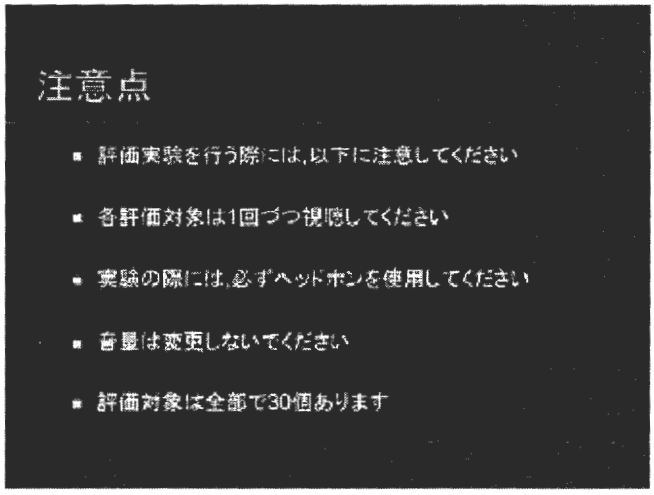

図20　評価実験を行う上での注意事項

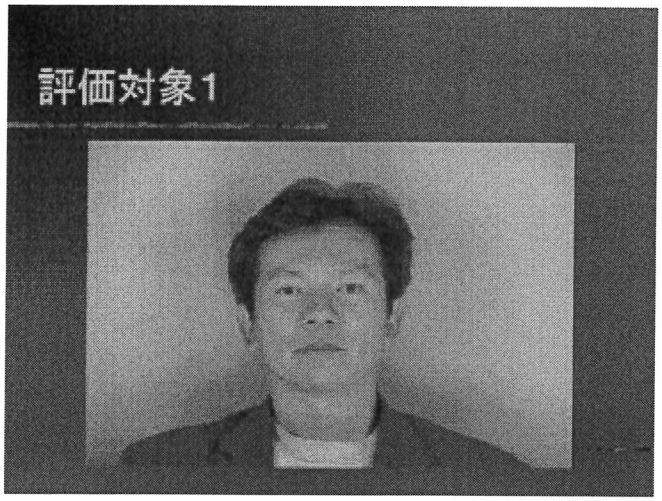

図21　評価対象の提示

6.3.2　考察

図22に，図18において被験者が，「非常によい，普通，悪い」と回答している5つの点（-462, -297, +66, +396, +495msec）における数字識別率を示す．

図22に着目すると，発話顔アニメーションよりも音声が早い方が，特に非同期の影響を受けやすく，自然発話の場合においてそれが顕著となった．これは合成発話の場合，それが合成されたものであると分かると無意識に音声の方に重点をおいて聞き取りを行ったためであると考えられる．逆に音声よりも口の動きの方が早い場合に着目すると，非同期の影響をにくいことが判明した．

以上の結果から，音声よりも口の動きを数10msec程度早くすることにより，自然な発話スタイルに近い発話顔アニメーションが合成できると考えられる．また，発話顔アニメーションを合成する場合に限らず，音声と顔画像を用いて音声認識を行う場合においても音声と発話顔アニメーションの間の非同期性を考慮に入れることにより，認識率が上がる可能性があるということが言える．

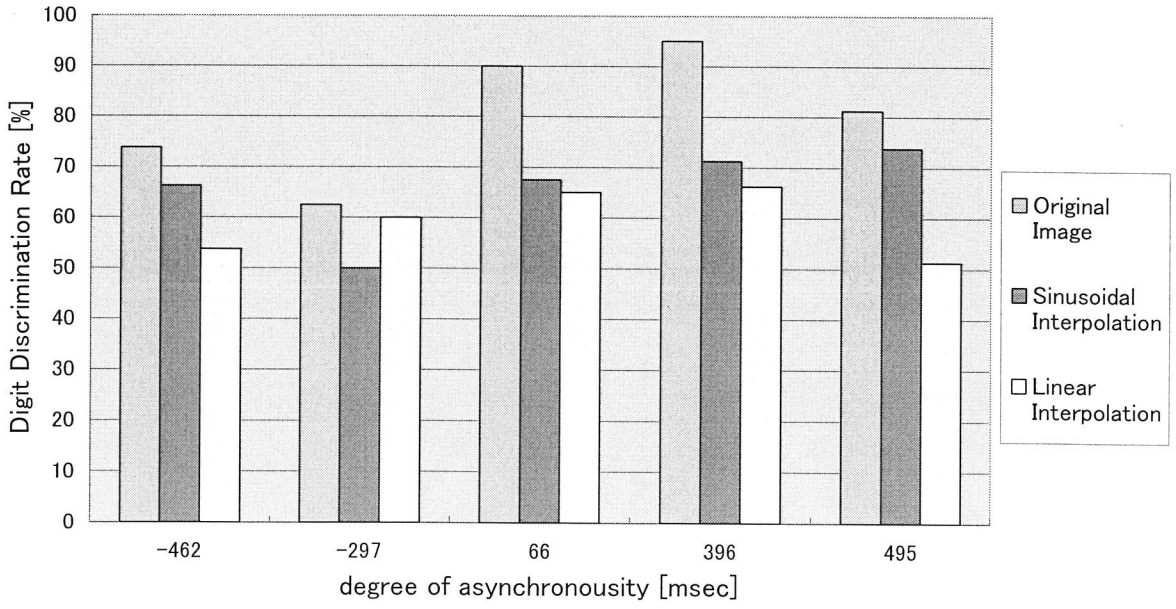

図22. 同期のずれが音声認識に与える影響

7. 合成発話顔アニメーションの品質評価

以上の考察から，合成発話顔アニメーションの品質を評価する．そして，提案手法である，2種類の口形状補間法のうち線形補間法と正弦波補間法のどちらが適切であるかを検討する．

線形補間法は，キーフレームのみ100%の口形を表すため，キーフレームがフレームレートの間隔と一致しない場合，基本口形を100%表すフレームが存在しないという問題点がある．このため，正弦波補間法の場合と比較して口の動きが読み取りづらいと感じたと考えられる．図9に着目すると線形補間法が正弦波補間法よりも，5～10%程度聞き取り率が下回っている．このため線形補間法よりも正弦波補間法の方が口形状の再現性が高くなっている．

一方で正弦波補間法は，基本口形を100%表すフレームを線形補間法よりも長く持続させることができるが，後続する口形への変化が線形補間と比較して急峻になる．しかしながら，人間の口の動きは，非線形な動きをするため，多くの被験者には正弦波補間法の方が自然であると感じたと考えられる．図10～11に着目すると顔画像の自然性，口の動きの滑らかさにおいて正弦波補間法が線形補間法よりも0.3～0.5点程度上回る結果となり，自然性．口の動きの滑らかさに関しても，正弦波補間法の方が高く，キーフレーム補間法として妥当であると考えられる．以上より，現状では正弦波補間法により作成した合成発話顔アニメーションが品質が高いと言える．

しかしながら，これらの評価における両者の差はわずかであり，自然発話顔アニメーションと同程度の発話顔が合成できていないというのが現状であるといえる．今後，より自然な発話顔アニメーションを合成する手法を検討する必要がある．

8. 結論

本稿では，雑音環境下において音声とともに発話顔アニメーションを被験者に提示し，音声の明瞭度や，発話顔アニメーションの視覚的な自然さ，音声と発話顔アニメーションの同期に対して検証することで，合成発話顔アニメーションの評価を行った．

評価結果より合成された発話口形の再現性は，未だに自然発話のものよりも低く(約50%)，自然性においてもやはり自然発話口形との差はあると言える．今後，再現性の高い発話口形を生成する手法を検討する必要がある．また，音声と発話顔アニメーションのとの同期について検証した結果，発話顔アニメーションを合成する場合には，音素の開始と同時に音素に対応する口の動きを始めるのではなく，数10msec程度口の動きを早くすることで，合成発話顔アニメーション自然さを改善できると考えられる．これは音声と発話顔画像を用いて音声認識を行う上でも同じことが言える．

ここで用いた評価方法は，主に著者等の提案手法のような客観評価の困難な個人性の除去された特徴から合成された発話顔アニメーションに対して有効であると考える．さらに，合成発話顔アニメーションの評価だけでなく，例えば顔画像の表情と音声のモダリティ間の同期が失われた場合，被験者の受ける印象はどのように変化するのか，また合成音声とともに顔画像を提示することにより，単にSNRでは表すことの出来ない合成音声の印象や明瞭性の評価等にも応用が可能であると考える．

9. 今後の課題

ここでの評価実験は，すべてマシン上で実施したものであり，評価対象を再生する際のマシンによる音声と映像間同期が取られていることを仮定している．しかしながら，実際には，音声と映像との間にごく小さい遅延があると考えられる．今後は，実験に使用するハードウェアにおける音声と映像との同期が正確に取られているのか，またどの程度の遅延があるのかについて調査する必要があると考えられる．また，これを検証した上で今回実施した実験を行う必要があると考えられる．

さらに，6.2節で示した音声と発話アニメーションとの非同期による影響について，さらに被験者を増やして分布を確かなものにすること，そして，本論文では考慮の対象としなかった音声と発話顔アニメーションとの音韻的な同期についても今後検証していく必要がある．

参考文献

[1]川本，下平，新田，西本，中村ら，”擬人化音声対話エージェントツールキットの基本設計”，情報処理学会研究報告 音声言語情報処理学会，2002-SLP-40-11, pp61-66, Feb 2002.

[2]H. McGurk and J. MacDonald, "Hearing lips and seeing voices", Nature 264, pp746-748 1976

[3]酒向，徳田，益子，小林，北村，”HMMに基づいた視聴覚テキスト音声合成－画像ベースのアプローチ－”，情報処理学会論文誌， Vol. 43. No7. pp0-7, 2002

[4]垣原，中村，鹿野，”HMMを用いた自然な発話動画像合成”，電子情報通信学会論文誌，Vol. J83-D-Ⅱ，No. 11, pp2498-2506, 2000

[5]T. Kuratate, H. Yehia, and E. Vatikiotis Bateson, “Kinematics-based Synthesis of Realistic Tracking Face “, Proc. International Conference on Auditory-Visual Speech, Processing, AVSP 98, pp. 185-190, 1998

[6]Hans Peter Graf, Eric Coatto, and Tony Ezzat, “Face Analysis for the Synthesis of Photo-Realistic Talking Heads”, Proc. 4th International Conference on Automatic Face and Gesture Recognition, pp. 189-194, 2000

[7]Eri Yamamoto, Satoshi Nakamura, and Kiyohiro Shikano, ” Subjective Evaluation for HMM-Based Speech-to-Lip Movement Synthesis”, AVSP98, pp 225-230, 1998

[8]Tony Ezzat, Gadi Geiger, and Tomaso Poggio, “Trainable Videorealistic Speech Animation”, ACM SIGGRAPH 2002, 2002

[9]緒方，森島，中村，”ビデオ翻訳システム－自動翻訳合成音声とモデルベースリップシンクシステムの実現－”，情報処理学会インタラクション， 2001

[10]伊藤，三澤，武藤， 森島，”仮想空間上におけるリアルな3次元口形状の生成”，電子情報通信学会2000年総合大会講演論文集，A-16-24, pp. 328, 2000

[11]T. Misawa, S. Nakamura, and S. Morishima, ” Automatic Face Tracking And Model Match-Move InVideo Sequence Using 3-D Face Model”, Proc. International Conference of Multi-Media, Expo, ICME 2001, 2001

[12]Shigeo Morishima, Satoshi Nakamura, "Multi-modal Translation System and its Evaluation", Proc. of ICMI’02, pp. 241-246, 2002

[13] S. Young et al, “The HTK Book”, Microsoft Corporation, 2000

[14]前島，森島，中村，”複数人話者会話シーンの動画像翻訳”， 電子情報通信学会技術研究報告，Vol. HCS2002-30, pp13-18, 2003