動画像情報の音声知覚への影響

3R－9

中園　薫

ＮＴＴソフトウェア研究所

１．はじめに

人間が音声知覚するとき、耳から聞こえる音だけでなく、視覚の影響も受けていることが知られている。たとえば、「ば」という音声と「が」と言っている口の動きの画像とを合わせて提示すると、他の音――たとえば「だ」――に知覚される[1]（McGurk 効果）。本稿では、音声知覚に影響を与えるのは動画像情報の中のどんな要素なのか、画像のフレームレートなどを変化させ、聴取実験をおこなうことによって議論する。

文献[2]において、フレームレートが変わると、McGurk 効果によって異聴が生じる度合が変わることを示した。しかし、フレームレートの変動によって生じるどのような視覚的刺激の変化要因が音声知覚に影響を与えたのか、真の原因を特定するまでには至らなかった。

ここでは、その要因として、（１）画像の動きがとびとびに不自然になることによって聴覚に与える影響を阻害する、（２）動画像のコマを間引くことによって音韻の決定をする上で重要な特定の視覚的刺激を持った画像（コマ）が落ちる、の２つを考える。この特定のコマとは、唇音と非唇音の間で異聴が顕著に見られることから、「唇を閉じた瞬間」の画像であると予想できる。

そこで、今回は、口を閉じた状態から「ば」と１回だけの発声したデータと、その前に「あ」の音をつけた発声したデータの２種類を用意した。（前者を "ba-type"、後者を "aba-type" と呼ぶ）そして、音声波形を見ながら、「ば」の音の立上りの瞬間からちょうど１秒前を開始点とし、そこから２秒間を刺激データの素材として切り出すことによって、刺激ごとの時間軸を正規化した。（図１）

この素材をもとに、30fps, 15fps, 10fps, 5fps, 3fps の刺激データをダウンサンプルした。これによって、どの刺激データも音の立上りの瞬間の画像を含むことになる。

さらに、aba-type の刺激については、フレームレートが十分低いときに図１に示したようにサンプルするフレームを半分ずらすと、口を閉じる瞬間がまったく入っていない刺激が作れる。（こうして作った刺激データを "5fps-Shift" と呼ぶ）

これらの刺激データを使って、提示、聞き取りの実験を行った。

２．実　験

[実験システム]

Macintosh Quadra 840 AV と DigitalFilm を利用した、デジタルビデオ環境による実験システムを構築した。詳しくは文献[2]を参照。

[刺 激]

"ba,ga", "pa,ka", "aba,aga", "apa,aka" の４組（それぞれのペアを類似刺激セットと呼ぶ）、８種類を 30 fps で録画した。これらをもとに、類似刺激のペアで音声と画像を入れ換えた刺激データを作成し、合計１６種類を基本刺激とした。ここで、音声を聞きにくくするために一定の白色雑音を加えた。さらに各刺激ごとに、"ba-type" については、30fps, 15fps, 10fps, 5fps, 3fps の５種類を、"aba-type" については、30fps, 15fps, 10fps, 5fps, 5fps-shift, 3fps, 3fps-shift の７種類のデータを作成した。

[被験者]

被験者は、正常な聴覚をもった日本人で、２０歳代半ば～３０歳代半ばの女性１０名を対象とした。

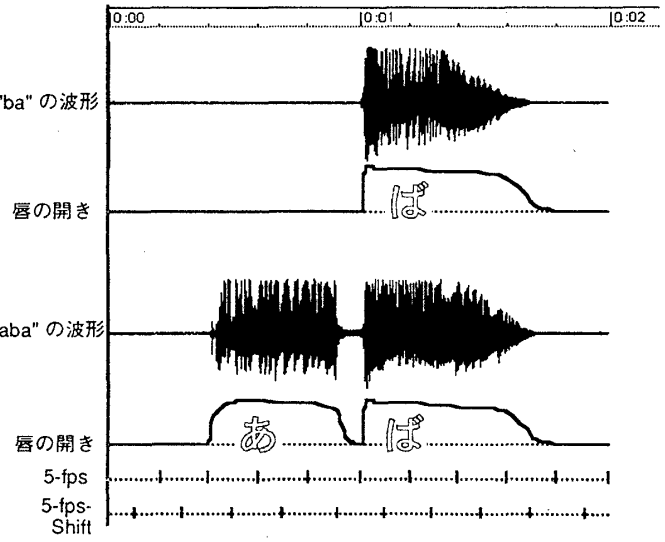

図１　音声波形とサンプリングのタイミング

Influence of visual information upon speech perception
Kaoru NAKAZONO
NTT Software Laboratories
3-9-11 Midori-cho Musashino-shi Tokyo, 180, Japan

［実験手続き］

被験者には、映像−−特に口の動き−−をよく見ながらどのように聞こえたか、３つの候補の中から答えるように指示した。３つの候補とは、"ba, ga" の場合 "ba, ga, da" , "pa, ka" の場合 "pa, ka, ta" , "aba, aga" の場合 "aba, aga, ada", "apa, aka" の場合 "apa, aka, ata" である。候補は提示刺激と同じディスプレー上に選択ボタンとして表示され、被験者はいちばん近いと感じたものをマウスクリックで選択する。

3． 結果と考察

図２に、ba-type の実験結果を、図３に aba-type の結果を示す。ここでは、音声刺激の通りに答えた場合を「正解」と呼び、正解率のみをグラフに表す。また、"ba" と "pa"、"ga" と "ka"、"da" と "ta" を合わせてカウントする。横軸はフレームレートである。ここで、提示した刺激を "ba/ba" という風に表記する時は、スラッシュ(" / ")の前が聴覚刺激の種類、後が視覚刺激の種類を表す。

図２から、"ba/ga" や "ga/ba" のように、音声と画像が矛盾する組合せにおいては、フレームレートが低くなるにつれて、正解率が上昇する傾向が見られる。（ここでは、"ba" や "pa" の音では必ず唇を閉じている瞬間が見えていることに注意）

図４に、aba-type の 5fps と 3fps においてサンプルタイミングをシフトさせた場合との比較を示す。5fps では、"aba/aba" は、シフトさせることによって正解率が大きく下がり、"aga/aba" は逆に正解率が上がっている。このことから、"aba" の "b" の破裂音のための唇を閉じる瞬間が見えることが、"ba" と "ga" を聞き分ける上で重要であることがわかる。ところがこの傾向は 3fps ではそれほど顕著に見られなくなる。これは、「動きのぎこちなさ」による妨害の方が大きくなるためであると考えられる。

さらに、フレームレートを落としていくと、唇音の場合、「ば」の破裂の瞬間が遅れて見えることに注意する。音声と画像との同期ずれによって、異聴の生じやすさが変化することが知られている[3]ので、図２、３でグラフの傾きが単調でない原因として、この同期ずれの問題も考えられる。

最後に、図２と図３を見較べると、"aba" の聴覚刺激＋"aga" の視覚刺激では、"ba" の 聴覚刺激＋"ga" の視覚刺激ほど異聴が生じていないことがわかる。子音に後続する母音の種類によってMcGurk効果の生じ方が異なることは知られている[4]が、子音に先行する母音の影響も受けることが示唆された。ただ、先行する母音の音声そのものか、画像か、いずれの影響によるものなのかは今後の課題である。

［参考文献］

1. McGurk, MacDonald, Nature 1976;264(Dec. 23/30):746-748.
2. 中園,信学技報 1994;(HC93-68)
3. 坂口,世木,出口, 音講論集 1993-5, 1-7-4
4. Greene, Kuhl & Melzoff, J.A.S.A. 1988;(84, S155)

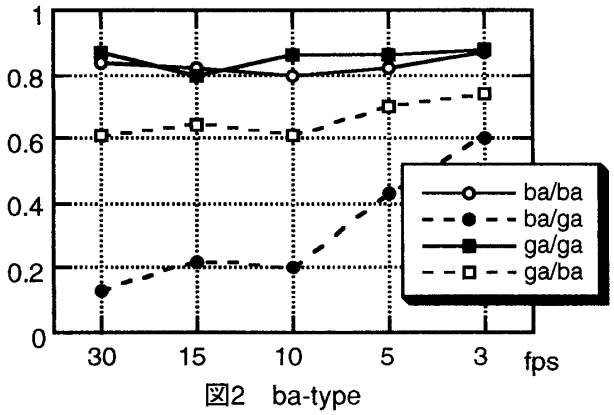

図2　ba-type

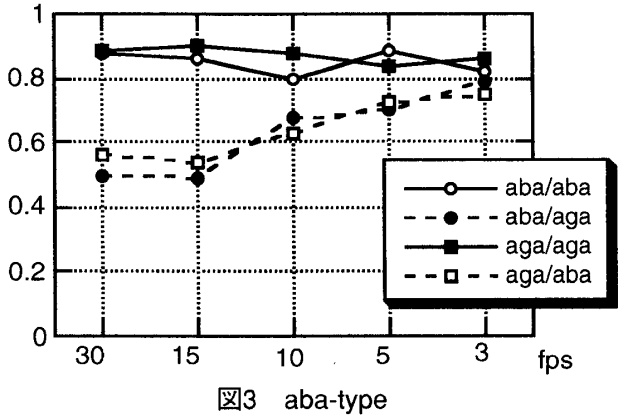

図3　aba-type

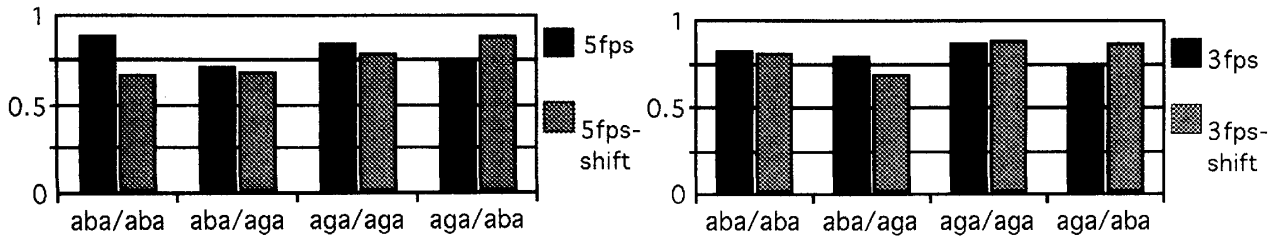

図4　サンプルタイミングシフトによる効果