Technical Report on Attention and Cognition (2006) No.13

映像と音声のタイミングおよび速度の違いが単語音声認知に及ぼす影響

田中 章浩 東京大学大学院人文社会系研究科／文学部
津村 光美 東北大学電気通信研究所／大学院情報科学研究科
坂本 修一 東北大学電気通信研究所／大学院情報科学研究科
鈴木 陽一 東北大学電気通信研究所／大学院情報科学研究科

We investigated the influence of timing shift and presentation rate difference between talker's moving image and voice on word intelligibility. We used 20 minimal pairs of 4-mora-word. Words were presented under visual-only, auditory-only, and auditory-visual (AV) conditions. Effect of AV asynchrony by time-expanded speech on AV benefit was compared with that by timing shift. Results showed that AV asynchrony by timing shift was recalibrated while AV asynchrony by time-expanded speech was not. These results suggest that recalibration of AV simultaneity requires constant timing difference between talker's moving image and voice.

Keywords: speech-rate conversion, audio-visual integration, audio-visual asynchrony, word recognition

問題・目的

話者の顔の映像を見ながら音声を聴取すると，映像がない場合よりも音声の了解度が向上することが，単音節から文章まで様々なレベルで確認されている．この場合，映像と音声は完全に同期している必要はなく，音声遅延が200 ms程度以内であれば映像の効果が得られる（McGrath & Summerfield, 1985など）．これは，視覚情報と聴覚情報の時間差を検出して補正するメカニズム（Fujisaki *et al*., 2004）に支えられている可能性がある．もしこの可能性が正しければ，映像と音声の時間差が一定でない状況下では映像の効果が低下することが予測される．そこで本研究では，話者映像と音声を刺激に用いて，映像と音声のタイミングおよび速度の違いが映像の効果に及ぼす影響について検討した．上記の可能性が正しければ，音声の提示タイミングを単に遅延させた場合，遅延量が音声区間全てにおいて一定のため，200 ms程度までは映像の効果は低下しないが，音声を時間伸長して映像と異なる速度にした場合には，後方に進むにつれて映像と音声の時間差が拡大するため，大幅に伸長すると補正メカニズムが機能せず，映像の効果が低下することが予測される．

方法

被験者　日本語を母語とし，正常な視力（矯正も含む），聴力をもつ大学生10名が実験に参加した．

刺激　三省堂「NTTデータベースシリーズ 日本語の語彙特性」から，1子音違い，異口形，音声単語親密度差0.875以下の4モーラ単語を20ペア選定した（Table 1に例を示す）．各モーラ位置において1子音違いの単語を5ペアずつ用いた．女性話者が単語を発話し，音声（話速平均6.9 mora/s，平均時間長583 ms）と映像を記録した．音声の加工にはSTRAIGHT（Kawahara *et al*., 1999）を利用した．

実験手続き　同じ速度の音声と映像を用い，音声を遅延して提示する条件（非同期AV：0，100，200，300，400 msの5水準．Figure 1(b)参照），音声に話速変換を施し，映像と音声の速度が異なる条件（伸長AV：0，100，200，300，400 ms伸長の5水準．Figure 1(a)参照），話速変換音声のみを提示する条件（伸長A：0，100，200，300，400 ms伸長の5水準），統制条件（統制A（原音声のみ提示），統制V（映像のみ提示））の計17条件を実施した．音声の提示レベルは60 dBA，S/Nは－10 dBで，DVデッキからアンプを通してスピーカ(B&W N-803)から提示した．映像は30 frame/sで提示した．回答用紙は，例えば「トウハツ」の場合は「ト」の部分が空欄になっており，空欄に入る1拍を記入させた．単語了解度は，空欄に記入された1拍の正答率とし，モーラ位置別に算出した．

Table 1. 実験で用いた単語の例．

モーラ位置	ペア単語の例	
1	トウハツ	モウハツ
2	ミガワリ	ミマワリ
3	トクダイ	トクバイ
4	ミズアゲ	ミズアメ

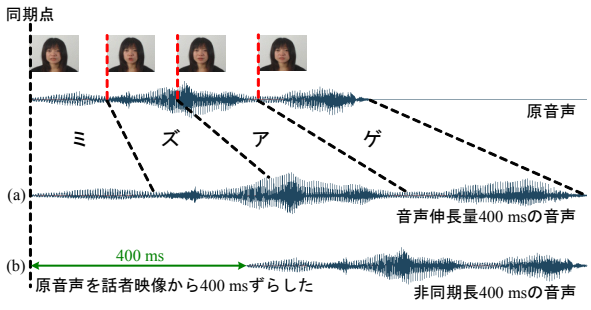

Figure 1. 映像提示と音声提示のタイムコース．

結果・考察

各条件，モーラ位置別の単語了解度をもとに，映像の効果の指標としてAV benefit [(*AV*－*A*)／(100－*A*)]を算出した．この指標はSumby & Pollack (1954) 以降多くの研究で採用されている指標であり，*A*は音声のみ提示する条件，*AV*は映像と音声を提示する条件の了解度を表す．非同期条件では統制A，伸長条件では伸長Aの各水準における了解度が*A* に該当する．非同期条件では非同期AVの各水準，伸長条件では伸長AVの各水準における了解度が*AV* に該当する．Figure 2に非同期条件, Figure 3に伸長条件でのAV benefitを示した．

Figure 2をみると，第1～3モーラでは非同期量の増大とともにAV benefitが低下する傾向が確認できる．ただし，非同期量が200 msを超えてもAV benefitは0にはなっていない．モーラ×非同期長の2要因分散分析の結果，交互作用($F(12,108) = 5.30, p < .01$)，モーラの主効果($F(3,27) = 20.87, p < .01$)，非同期長の主効果($F(4,36) = 10.08, p < .01$)が有意であった．モーラ位置別に，非同期長0 msと他の非同期長のAV benefitの差をダネット法によって比較した結果，第1，第3モーラでは非同期長400 msのAV benefitが，第2モーラでは非同期長300 ms，400 msのAV benefitが，それぞれ有意に低かった．第4モーラでは，非同期長0 msと他の非同期長のAV benefitには有意差が認められなかった．

非同期条件では，どのモーラ位置でも映像と音声の時間差は一定である．したがって，もし単語認知において前後のモーラの情報を利用せずに視聴覚統合が生じるのであれば，非同期条件におけるAV benefitはどのモーラ位置でも一定となるはずである．しかし，非同期条件におけるAV benefitは後方のモーラに進むにつれて上昇した．この結果は，視聴覚統合の過程で前のモーラの情報が利用されることを示唆している．

非同期条件では第1～3モーラでは非同期量の増加とともにAV benefitが低下したのに対し，第4モーラでは非同期量によらずAV benefitは一定であった．この結果は，第1～3モーラが提示される間に映像と音声の時間差を検出し，第4モーラ提示の段階では時間差を補正するメカニズムが機能したためと解釈できる．

一方，Figure 3をみると，第1モーラでは伸長量によらずAV benefitは一定であるが，後方のモーラでは伸長とともにAV benefitが低下する傾向が確認できる．

モーラ×非同期長の2要因分散分析の結果，交互作用($F(12,108) = 1.92, p < .05$)，モーラの主効果($F(3,27) = 14.68, p < .01$)，音声伸長量の主効果($F(4,36) = 3.03, p < .05$)が有意であった．モーラ位置別に，非同期長0 msと他の非同期長のAV benefitの差をダネット法によって比較した結果，第1～3モーラでは有意差がなく，第4モーラでは，音声伸長量200 ms以上で有意にAV benefitが低下した．この結果は，映像と音声の時間差が一定の場合（非同期条件）には補正メカニズムが機能する一方，速度が異なり時間差が一定でない場合（伸長条件）には補正メカニズムが機能せず，後方のモーラではAV benefitが低下した可能性を示唆している．

謝辞

STRAIGHTの利用を認めて下さった，和歌山大学河原英紀教授に深く感謝する．本研究についてご議論いただいたNHK放送技術研究所の皆様に感謝申し上げる．本研究は，東北大学電気通信研究所共同プロジェクト研究(H17/A12)の一環としておこなった．

引用文献

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. (2004). Recalibration of audio-visual simultaneity. *Nature Neuroscience*, 7, 773-778.

Kawahara, H., Masuda-Katsuse, I., & de Cheveigne, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27, 187-207.

McGrath, M. & Summerfield, A.Q. (1985). Intermodal timing relations and audio-visual speech. recognition by normal-hearing adults. *Journal of the Acoustical Society of America*, 77, 678-685.

Sumby, W.H. & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212-215.

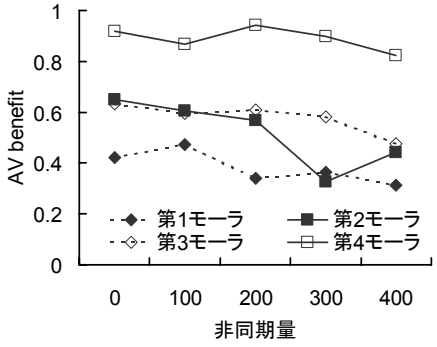

Figure 2. 非同期条件におけるモーラ位置別の AV benefit.

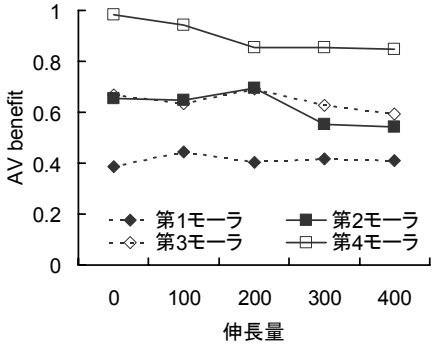

Figure 3. 伸長条件におけるモーラ位置別の AV benefit.