Auditory-Visual Perception of Prosodic Information: Inter-linguistic Analysis - Contrastive Focus in French and Japanese

Marion Dohen, Chun-Huei Wu, Harold Hill

Auditory-Visual Perception of Prosodic Information: Inter-linguistic Analysis – Contrastive Focus in French and Japanese

Marion Dohen[1,2], *Chun-Huei Wu*[1], *Harold Hill* [2,3]

[1] Speech and Cognition Department, ICP, GIPSA-lab, France
[2] ATR, Cognitive Information Science Laboratories, Japan
[3] School of Psychology, University of Wollongong, Australia
Marion.Dohen@gipsa-lab.inpg.fr

Abstract

Audition and vision are combined for the perception of speech segments and recent studies have shown that this is also the case for some types of supra-segmental information such as prosodic focus. The integration of vision and audition for the perception of speech segments however seems to be less important in Japanese. This study aims at comparing auditory-visual perception of prosodic contrastive focus in French and in Japanese. Two parallel focus identification tests were conducted for three modalities: AV, A and V in the two languages. Four speakers were recorded in both languages. For French, there was no AV advantage due to a ceiling effect. The same was observed for Japanese even though auditory only performances did not reach a ceiling. The results suggest that there are visual cues to prosodic focus in Japanese as well as in French but that these are not systematically combined with auditory information to enhance AV perception in Japanese. However, we also found that in some cases, especially when auditory alone perception is poor, visual and auditory information can be combined to enhance perception in Japanese.

Index Terms: Auditory-visual perception, prosody, contrastive focus, French, Japanese, inter-linguistic differences

1. Introduction

Vision participates to overall perception of speech segments. When the acoustic information is degraded, adding vision enhances auditory alone perception in many languages: for example for the perception of speech in noise ([1-9]). This auditory-visual enhancement is also measured for clear speech: when it is produced by a non-native speaker, when it is produced in a foreign language or when it is semantically complex ([10]). Moreover, the influence of vision on overall perception of speech segments is also put forward when the visual and auditory information are incongruent as in the McGurk effect (an auditory [ba] dubbed to a visual [ga] results in a [da] percept: [11]). This concerns the segmental perception of speech for which it was shown that auditory and visual information are not simply superimposed and redundant but are rather integrated and complementary for speech perception.

Supra-segmental aspects of speech (prosody) have been mainly conceived of acoustic/auditory. Prosodic contrastive focus is used to emphasize a word or a group of words in an utterance (*e.g.*, "SARAH ate the apple" as opposed to "Thomas ate the apple"). A number of studies have shown that there are potentially visible correlates to prosodic focus ([12-25]). Moreover, these correlates are actually used and visual only perception of prosodic focus is possible ([26-29, 22, 24]). [30] analyzed the relative cue value of different facial correlates to prosodic focus. [31] showed that, when the acoustic prosodic information is degraded (whispered speech), it appears that audition and vision are integrated in bimodal perception of prosodic focus in French. It therefore appears that vision also plays a role in the perception of supra-segmental information.

It seems that the integration of auditory and visual information during segmental perception of speech varies across languages. For example, the McGurk effect is weaker for Japanese speakers than for English speakers ([32]) and probably French speakers. It was also shown that, whereas multimodal integration skills develop with age for English speakers, this is not the case for Japanese speakers ([33]). This suggests that visual information plays a less important role in overall perception of speech segments in Japanese than in English (and probably also in French). One can wonder whether this is also the case for the perception of supra-segmentals.

The aim of this study is therefore to compare auditory-visual perception of prosodic information (contrastive focus) in French and in Japanese.

2. Experimental Methods

Two parallel perception experiments using exactly the same paradigm were conducted respectively for French and for Japanese.

2.1. Stimuli

2.1.1. Corpora

Two five-sentence corpora were designed. Corpus 1 was a French corpus consisting of subject-verb-object (SVO) sentences. Corpus 2 was a Japanese corpus consisting of SOV sentences. The two corpora can be found in appendices 1 and 2.

2.1.2. Audio-visual recordings

Both corpora were recorded for four native speakers (French: Sf1, Sf2, Sf3, Sf4; Japanese: Sj1, Sj2, Sj3, Sj4). The video recordings were made in parallel to motion capture recordings using optotrak (IRed facial markers: see [31] for a discussion on the influence of facial markers on perception). The recordings were conducted in a sound attenuated room at the ATR Cognitive Information Science Laboratories. In both languages, three focus conditions were recorded: neutral, subject focus (SF) and object focus (OF). A correction task was used in order to trigger focus in the most natural way

possible (speakers were not directly asked to produce focus). The speakers listened to a prompt in which two people (S1 and S2) were talking. S1 first pronounced a sentence from the corpus (corpus 1 for French and corpus 2 for Japanese) which S2 then repeated in a question mode because he/she was not sure to have understood correctly one of the constituents from the sentence (S or O). The recorded speaker then had to correct S2 and thus produced contrastive focus on the mispronounced constituent. The recording therefore went as follows (capital letters signal focus; example provided for French but similar procedure used for Japanese):

Audio prompt:

S1: Lou mima le lama.'Lou mimed the lama.'

S2: S1 said: Jo mima le lama? 'S1 said: Jo mimed the lama?'

Speaker utters:

LOU mima le lama.

No indication was given to the speakers on how to produce focus (*e.g.*, which syllable(s) was(were) to be focused). When S2 had correctly understood (he/she produced the correct sentence in a question mode), the recorded speaker was instructed to produce a neutral version (broad focus) of the sentence *i.e.* without focusing any particular constituent. An example of a recorded image is provided in Figure 1.

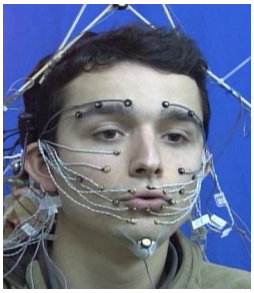

Figure 1: *Image extracted from the videos recorded.*

2.2. Experimental paradigm

The tests took place in a quiet room in which the participants were isolated both from outside noise and from the experimenters (French experiment: Speech and Cognition Department, GIPSA-lab; Japanese experiment: ATR Cognitive Information Science Laboratories). The videos were shown on a video monitor. The speaker's head on the screen was approximately life size. Two loudspeakers were located on the screen's sides.

Participants were told that they would be following part of a conversation between two people (S1 and S2). S1 would first utter a sentence from the corpus. Not having heard the sentence very well, S2 would question S1 by repeating the sentence he had understood, in a question mode. S1 would then repeat the first sentence he had uttered correcting the constituent (S or O) that S2 had misunderstood. He would therefore insist on this particular constituent (*i.e.* focus it). Participants were told that they would neither hear nor see S1's first utterance as well as S2's. They would either see only (V), or hear only (A), or hear and see (AV) S1's correction. Below is an example of how the experiment went:

S1 (participants do not hear nor see):

Lou mima le lama.

S2 (participants do not hear nor see):

Jo mima le lama?

S1 (participants hear or see or hear and see):

LOU mima le lama.

Participants were told that, in some cases, no correction would be performed by S1 (S2 had correctly understood). In that case, S1 would simply repeat the initial sentence (*i.e.* neutral version of the sentence). The task was to identify which constituent (S, O or none) had been misunderstood by S2 and thus corrected by S1. Participants were asked to highlight the constituent they had identified as being corrected on an answer sheet such as the one presented below (blank column on the right for 'no correction' responses):

Lou	le lama.	

Participants were thus indirectly asked to identify whether a constituent in the utterance had been focused and which one. They were never told about "focus" or about the experiment's aim.

Six movie clips were elaborated: 2 random combinations of the stimuli for each modality (AV, A, V). Each participant evaluated all the stimuli in all modalities. The order in which the modalities were evaluated was varied across participants. Before taking the test in each modality, the participants went through a short training.

A total of 60 stimuli were evaluated by each participant (5 sentences, 3 focus conditions, 4 speakers) for each modality (A, V and AV). This represents a total of 180 stimuli.

2.3. Participants

Sixteen native speakers of French (8 men and 8 women) and thirteen native speakers of Japanese respectively participated in each of the two experiments. All the participants reported normal or corrected to normal vision and no hearing problems.

3. Results

3.1. General results

Table 1 provides the percentages of correct answers (focus condition identified correctly) for each modality and for both experiments. Both experiments had a 3×4×3 design with the following within subject factors: modality (3 levels: AV, A and V), speaker (4 levels) and focus condition (3 levels: neutral, SF and OF). Three-way repeated analyses of variance were conducted for each experiment on the percentages of correct answers with the above within subject factors as independent factors. The sphericity of the data was checked for using the Mauchly sphericity test. When the test was significant, the Huynh-Feldt correction was applied on the number of degrees of freedom and all the results presented below correspond to these corrected results (when necessary). For the sake of clarity, even when the results were corrected, they will be reported with the "true" numbers of degrees of freedom. The results of pairwise comparisons were corrected for using the Bonferroni correction.

Table 1. *Mean percentages of correct answers across all subjects for each modality and for the two experiments (chance level: 33.3%).*

modality	French	Japanese
	% correct	% correct
AV	92.9	68.6
A	92.5	67.2
V	66.4	44

3.1.1. French

It appears that, for all modalities, the percentages of correct answers were significantly above chance (33.3%) suggesting that it is possible to identify focus conditions whatever the modality. The statistical analysis revealed significant main effects of modality ($F(2,30)=144.637$; $p<.001$) and speaker ($F(3,45)=59.131$; $p<.001$). There was no significant main effect of focus condition ($F(2,30)=2.5$; $p=.123$). The significant interactions will be discussed in the detailed analysis (section 3.2). The significant main effect of modality illustrated the fact that, for French, the results in the A and AV modalities were significantly better than those in the V modality (ANOVA contrast: $p<.001$). The results in the AV and A conditions were not significantly different (ANOVA contrast: $p=.806$). Therefore, no AV advantage was measured (AV-A~0). This can be explained by a ceiling effect: the A performances were close to perfect (92.5%) and no improvement was possible. The significant effect of speaker illustrates the fact that overall perception was the best for Sf1 and the worst for Sf3.

3.1.2. Japanese

It appears that, for all modalities, the percentages of correct answers were significantly above chance (33.3%) suggesting that it is possible to identify focus conditions whatever the modality. The statistical analysis revealed significant main effects of modality ($F(2,24)=125.518$; $p<.001$) and speaker ($F(3,36)=8.361$; $p<.001$). There was no significant main effect of focus condition ($F(2,24)=2.437$; $p=.109$). The significant interactions will be discussed in the detailed analysis (section 3.2). The significant main effect of modality illustrated the fact that, for Japanese, the results in the A and AV modalities were significantly better than those in the V modality (ANOVA contrast: $p<.001$). The results in the AV and A conditions were not significantly different (ANOVA contrast: $p=.425$). Therefore, no AV advantage was measured (AV-A~0). In this case, it cannot be explained by a ceiling effect since the A only performances were not very high and far from perfect (67.2%). The significant effect of speaker illustrates the fact that overall perception was the best for Sj1 and the worst for Sj3. Overall, the results corresponding to Sj1 and Sj4 were significantly better than those corresponding to Sj2 and Sj3.

3.2. Detailed analysis

Figure 2 provides the percentages of correct answers and standard errors for each speaker and each modality.

3.2.1. French

Figure 2 shows that, overall, the same general pattern (AV~A>V) was observed for all speakers. The statistical analysis revealed a significant interaction between modality and speaker ($F(6,90)=20.242$; $p<.001$). This illustrates the fact that the difference between the V score and the AV and A scores was much more important for speaker Sf3 than for the other speakers. There was a significant interaction between modality and focus condition ($F(4,60)=3.858$; $p=.007$). This illustrates the fact that, for the neutral focus condition, there was less difference between A (or AV) and V than for the other focus conditions. There was also a significant interaction between speaker and focus condition ($F(6,90)=18.824$; $p<.001$). This illustrates the fact that for Sf1 and Sf4, the results in SF and OF were better than those in the neutral condition whereas for Sf2 and Sf3, the results in SF and neutral were better (much better for Sf3) than those for OF. On the whole, SF was easier to detect for all speakers.

3.2.2. Japanese

There was a slightly significant interaction between modality and speaker ($F(6,72)=2.828$; $p=.016$). This illustrates the fact that there was a much larger difference between A (or AV) and V performances for Sj1 and Sj3 than for the other speakers. The A results were the best for Sj1 and the worst for Sj2 and Sj4 (see Figure 2). The AV results were the best for Sj1 and grouped for Sj4, Sj2 and Sj3 (see Figure 2). The V results were the best for Sj4 and the worst for Sj3 (close to chance level; see Figure 2). There was a significant interaction between modality and focus condition ($F(4,48)=8.605$; $p<.001$). This illustrates the fact that for the neutral focus condition, there was less difference between A (or AV) and V than for the other focus conditions. There was also a significant interaction between speaker and focus condition ($F(6,72)=12.907$;$p<.001$). This illustrates the fact that for Sj3, the performances were the worst in the neutral condition whereas they were the best for Sj1 and Sj3 in that condition. For all speakers except Sj3, performances were better in the neutral condition than in the OF condition.

An inter-speaker analysis showed that an AV advantage was actually measured for speakers Sj4 and Sj2 (see Figure 2). An inter-stimulus analysis was also conducted. The mean percentages of correct answers were computed for each stimulus over all the participants. This showed that the fact that no general AV advantage was measured (AV-A=0) actually corresponded to a mean null average of stimuli for which there was an AV advantage (AV-A>0) and stimuli for which the AV score was actually lower that the A score (AV-A<0). We therefore analyzed the A and V scores of stimuli corresponding to the following cases:

1. AV disadvantage (AV-A<0);
2. no AV advantage (AV-A=0) ;
3. AV advantage (AV-A>0).

For case 1, it appears that V only performances were equal to or lower than chance suggesting that, either there was no visual information (V score equal to chance) or that this information could be misleading (V score lower than chance). This latter case could explain why the A score is lower and not equal to the AV score. For case 2, the V score was higher than chance suggesting that there was some visual information but that it did not seem to be combined to auditory information to enhance perception. It may have been redundant information. For case 3, the V only score was higher than chance and the A score was not very high. The

resulting AV score was higher than both the A and V scores. In this case, it therefore appears that auditory and visual information were combined to enhance perception.

4. Discussion & Conclusion

The aim of this study was to compare auditory-visual perception of a prosodic feature (prosodic contrastive focus) in French and in Japanese. Previous studies have indeed showed that there are differences in the integration of audition and vision for the perception of speech segments between languages. In particular, it seems that the visual information is used less or not at all in Japanese (see [32]). Two parallel experiments were conducted for French and Japanese using exactly the same paradigm to test the auditory-visual perception of prosodic contrastive focus *i.e.* supra-segmental information. Participants were indirectly asked to identify focus conditions in three modalities: Auditory-visual (AV), Auditory only (A) or Visual only (V). For both languages, the productions of four different speakers were evaluated in order to study inter-speaker variations. As expected, for French, AV and A performances were close to 100% and no improvement was measured from A to AV due to a ceiling effect. V performances were significantly higher than chance suggesting that there was visual information and that it could be perceived (confirmation of previous studies). Inter-speaker analyses suggested the V performances depended on the speaker. A previous study ([31]) using a whispered speech paradigm had shown that, when the acoustic prosodic information is degraded (no ceiling effect), the visual information is combined to the auditory information to enhance perception. Further analyses had suggested that the auditory and the visual information are complementary (rather than redundant) and integrated (rather than superimposed).

Surprisingly, in Japanese, A performances were well below 100% making AV improvement possible (no ceiling effect). AV performances were however not significantly different from A performances. V performances were significantly above chance suggesting that there was visual information and that it could be perceived. However, it did not seem to be combined to auditory information in auditory-visual perception. An AV advantage was however measured for two speakers. What is interesting is that the greatest AV advantage was measured for the speaker for which the A performances were the poorest and the V performances, the best. This suggests that Japanese perceivers can combine auditory and visual information to enhance perception especially when A perception is really poor. Moreover an inter-stimulus analysis showed that, for the stimuli for which an AV advantage was measured, the resulting AV score was higher than both the A and V scores taken separately which are not very high. This therefore suggests that the visual information may be less systematically used in Japanese but that when it is truly necessary (A only perception low), it can be combined to auditory information to enhance general perception. These results are not contradictory with those from [32] which showed that the McGurk effect was weaker in Japanese. It is indeed possible that the Japanese speakers perceive the visual information but that it is less systematically integrated. In this case, in the McGurk effect for which the auditory stream is clear (no noise), the perceivers would not integrate the auditory and the visual information. In this case, the visual information would simply not be used. The cognitive processes involved however still remain unclear and need further investigation.

Another possible explanation for the results of the Japanese experiment could be that the stimuli used were not good enough. The recording method and elicitation procedure used were however the same for both experiments. It could also be possible that the production of purely prosodic focus (without syntactic marking) is not as natural in Japanese as it is in French. This is the reason why we would like to explore the auditory-visual perception of other prosodic features in future work such as interrogation for example.

Acknowledgements: We thank our 8 very patient speakers as well as all the participants who took part in the test.

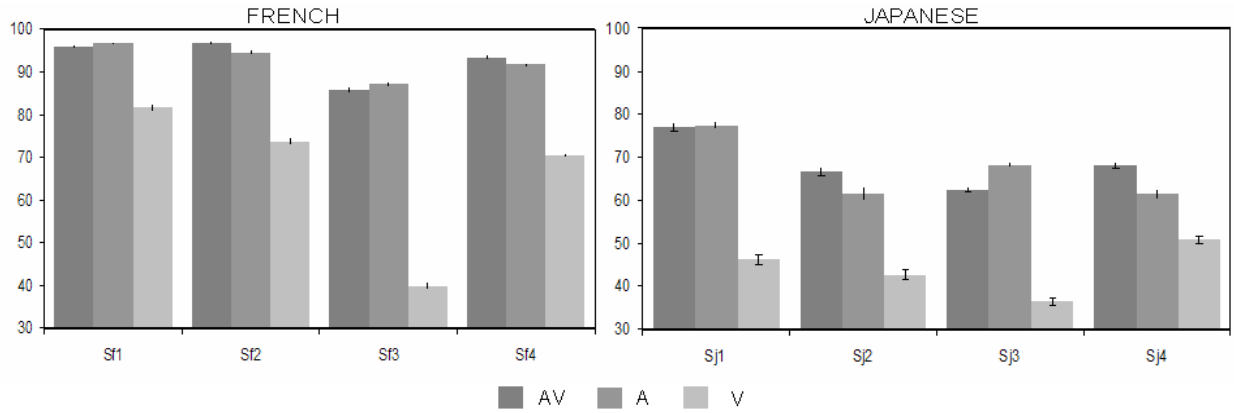

Figure 2: *Percentages of correct answers and standard errors for each speaker and each modality for both experiments (French and Japanese).*

5. References

[1] Sumby, W. H. and Pollack, I., "Visual contribution to Speech Intelligibility in Noise", *J. Acoust. Soc. Amer.*, 26(2): 212-215, 1954.

[2] Miller, G. A. and Nicely, P., "An Analysis of Perceptual Confusions among some English Consonants", *J. Acoust. Soc. Amer.*, 27(2): 338-352, 1955.

[3] Neely, K. K., "Effects of visual factors on the intelligibility of speech", *J. Acoust. Soc. Amer.*, 28(6): 1275-1277, 1956.

[4] Erber, N. P., "Auditory-visual perception of speech", *J. Spe. Hear. Dis.*, 40(4): 481-492, 1975.

[5] Binnie, C. A., Montgomery, A. A., and Jackson, P. L., "Auditory and visual contributions to the perception of consonants", *J. Spe. Hear. Res.*, 17(4): 619-630, 1974.
[6] Summerfield, A. Q., "Use of visual information for phonetic perception", *Phonetica*, 36: 314-331, 1979.
[7] MacLeod, A. and Summerfield, A. Q., "Quantifying the contribution of vision to speech perception in noise", *Brit. J. Audiol.*, 21: 131-141, 1987.
[8] Grant, K. W. and Braida L. D., "Evaluating the Articulation Index for audiovisual input", *J. Acoust. Soc. Amer.*, 89: 2952-2960, 1991.
[9] Benoît, C., Mohamadi, T., and Kandel. S., "Effects of Phonetic Context on Audio-Visual Intelligibility of French", *J. Spe. Hear. Res.*, 37: 1195-1203, 1994.
[10] Reisberg, D., McLean, J., and Goldfield, A., "Easy to Hear but Hard to Understand: A Lip-reading Advantage with Intact Auditory Stimuli", In Dodd, B. and Campbell, R. (Eds.), *Hearing by eye: The psychology of lip-reading*, Lawrence Erlbaum Associates, Hillsdale (USA), p. 97-114, 1987.
[11] McGurk, H. and MacDonald, J., "Hearing lips and seeing voices", *Nature*, 264: 746-748, 1976.
[12] Kelso, J. A. S., Vatikiotis-Bateson, E., Saltzman, E., and Kay, B. A., "A qualitative dynamic analysis of reiterant speech production: phase portraits, kinematics, and dynamic modelling", *J. Acoust. Soc. Am.*,, 77(1): 266-280, 1985.
[13] Summers, W. V., "Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses", *J. Acout. Soc. Am.*, 82(3): 847-863, 1987.
[14] Vatikiotis-Bateson, E. and Kelso, J. A. S., "Rhythm type and articulatory dynamics in English, French and Japanese", *J. Phonetics*, 21: 231-265, 1993.
[15] De Jong, K., "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation", *J. Acoust. Soc. Am.*, 97(1): 491-504, 1995.
[16] Harrington, J., Fletcher, J., and Roberts, C., "Coarticulation and the accented/unaccented distinction: evidence from jaw movement data", *J. Phonetics*, 23: 305-322, 1995.
[17] Loevenbruck, H., "An investigation of articulatory correlates of the Accentual Phrase in French", *Proceedings of the 14th ICPhS*, vol. 1, pp. 667-670, 1999.
[18] Erickson, D., Maekawa, K., Hashi, M., and Dang, J., "Some articulatory and acoustic changes associated with emphasis in spoken English", *Proceedings of ICSLP 2000*, vol. 3, pp. 247-250, 2000.
[19] Loevenbruck, H., "Effets articulatoires de l'emphase contrastive sur la Phrase Accentuelle en français", *Proceedings of XXIIIrd Journées d'Etude sur la Parole JEP 2000*, pp. 165-168, 2000.
[20] Erickson, D., "Articulation of Extreme Formant Patterns for Emphasized Vowels", *Phonetica*, 59, 134-149, 2002.
[21] Keating, P., Baroni, M., Mattys, S., Scarborough, R., Alwan, A., Auer, E. T., and Bernstein, L. E., "Optical Phonetics and Visual Perception of Lexical and Phrasal Stress in English", *Proceedings of ICPhS 2003*, pp. 2071-2074, 2003.
[22] Dohen, M., Lœvenbruck, H., Cathiard, M.-A., and Schwartz, J.-L., "Visual perception of contrastive focus in reiterant French speech", *Speech Communication*, 44: 155-172, 2004.
[23] Cho, T., "Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /a,i/ in English", *J. Acoust. Soc. Am.*, 117(6): 3867-3878, 2005.
[24] Dohen, M. and Lœvenbruck H., "Audiovisual Production and Perception of Contrastive Focus in French: a multispeaker study", *Proceedings of Interspeech/Eurospeech 2005*, pp. 2413-2416, 2005.
[25] Beskow, J., Granström, B., and House, D., "Visual correlates to prominence in several expressive modes", *Proceedings of Interspeech 2006 – ICSLP*, pp. 1272-1275, 2006.
[26] Thompson, D. M., "On the detection of emphasis in spoken sentences by means of visual, tactual, and visual-tactual cues", *J. Gen. Psychol.*, 11: 160-172, 1934.
[27] Risberg, A. and Agelfors, E., "On the identification of intonation contours by hearing impaired listeners", *Speech Transmission Laboratory - Quarterly Progress Report and Status Report*, 19(2-3): 51-61, 1978.
[28] Risberg, A. and Lubker, J., "Prosody and speechreading", *Speech Transmission Laboratory Quarterly Progress Report and Status Report*, 19(4): 1-16, 1978.
[29] Bernstein, L. E, Eberhardt, S. P., and Demorest, M. E., "Single-channel vibrotactile supplements to visual perception of intonation and stress", *J. Acoust. Soc. Am.*, 85(1): 397-405, 1989.
[30] Swerts, M. and Krahmer, E., "Facial expression and prosodic prominence: Effects of modality and facial area", *J. Phonetics*, 36(2): 219-238, 2008.
[31] Dohen, M. and Loevenbruck, H., "Interaction of audition and vision for the perception of prosodic contrastive focus", *Language & Speech*, In Press.
[32] Sekiyama, K. and Tohkura, Y., "Inter-language differences in the influence of visual cues in speech perception", *J. Phonetics*, 21: 427–444, 1993.
[33] Sekiyama, K. and Burnham, D., "Issues in the development of auditory-visual speech perception: Adults, infants and children", *Proceedings of Interspeech 2004 ICSLP*, pp. 1137-1140, 2004.

6. Appendices

6.1. Appendix 1: Corpus 1

Lou mima le lama.
'Lou mimed the lama.'
Le nominé lu les longs mots.
'The nominated read the long words.'
La nounou vit Lou.
'The nanny saw Lou.'
Les loups mimaient Marilou.
'The wolves mimed Marilou.'
Lou ramena Manu.
'Lou gave a lift back to Manu.'

6.2. Appendix 2: Corpus 2

まゆみ　は　りんご　を　食べます。
'Mayumi eats the apple.'
たかし　は　かびん　を　つくりました。
'Takashi made a vase.'
はるえ　は　いぬ　を　描きます。
'Harue draws a dog.'
みほ　は　晴れの日　を好みます。
'Miho prefers a sunny day.'
ひろゆき　は　やま　を　登りました。
'Hiroyuki climbed the mountain.'