

Comparison of audiovisual temporal synchrony perception with visual motion perception suggests a general feature-matching model for cross-attribute binding

Shin'ya Nishida and Waka Fujisaki

NTT Communication Science Laboratories, NTT Corporation, Japan

INTRODUCTION

Temporal synchrony is a critical cue for binding audiovisual information. For understanding of the sensory mechanism responsible for audiovisual synchrony detection, we compared its properties with those of visual motion detection, since extensive investigation has been made on this computationally analogous task—either task bears the correspondence problem (Marr, 1982). Previous studies have revealed that at least two mechanisms contribute to visual motion detection. One is low-level “motion-energy computation” that is fast and pre-attentive, while the other is high-level “feature tracking” that is slow and attentive: Long-range vs. Short-range (Braddick, 1974, *Vision Res*); Passive vs. Active (Cavanagh, 1992, *Science*); 1st-order vs. 3rd-order (Lu & Sperling, 1995, *Nature*). Here we show that the sensory mechanism responsible for audiovisual synchrony detection has a strong resemblance to feature tracking motion mechanism (see Fig. 1.)

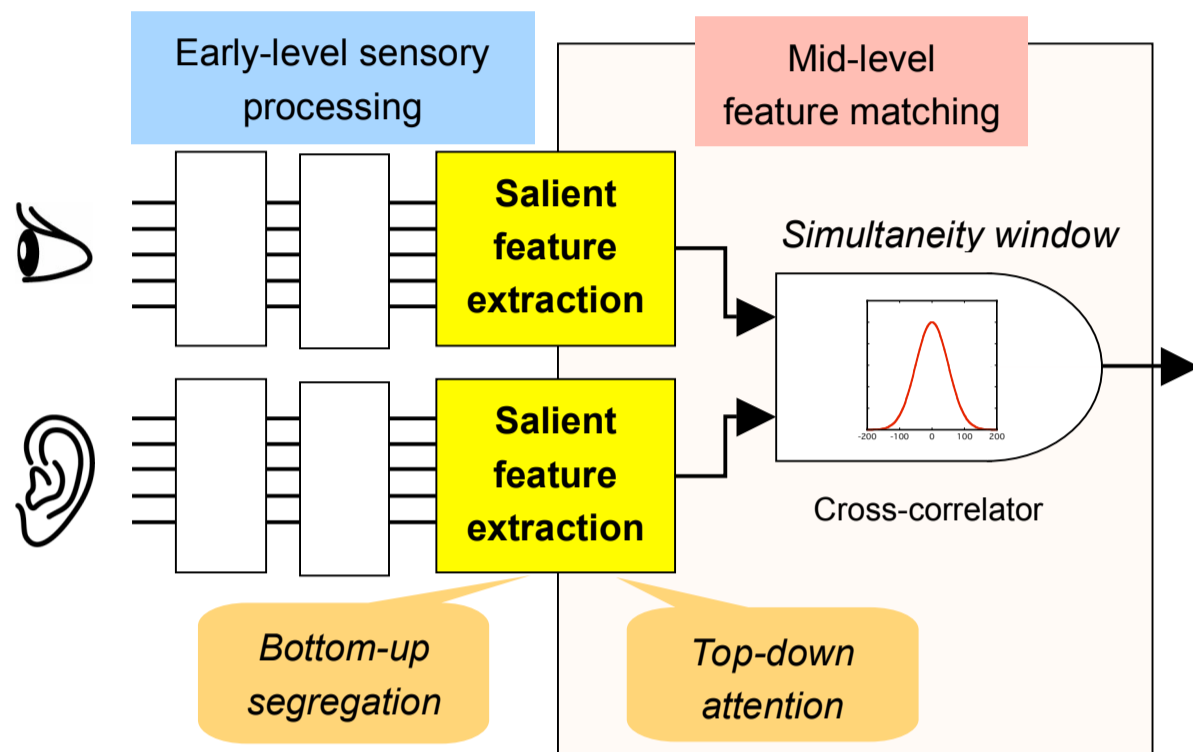


Fig. 1: Our model of audiovisual synchrony detection based on salient feature matching, whose structure is analogous to the models proposed for feature tracking motion mechanism.

TEMPORAL RESOLUTION

The upper temporal limit of feature tracking motion perception is fairly low (3-6 Hz) (Lu & Sperling, 1995, *Vision Res*). Similarly, the temporal limit of detecting audiovisual synchrony for repetitive pulse train is only ~4Hz (Fig. 2) (Fujisaki and Nishida, 2005, *Exp Brain Res*). In addition, difficulty of detecting audiovisual synchrony for high-density random-pulse train (Fig. 3, Black line) suggests that there is no “motion-energy computation” for cross-modal synchrony perception.

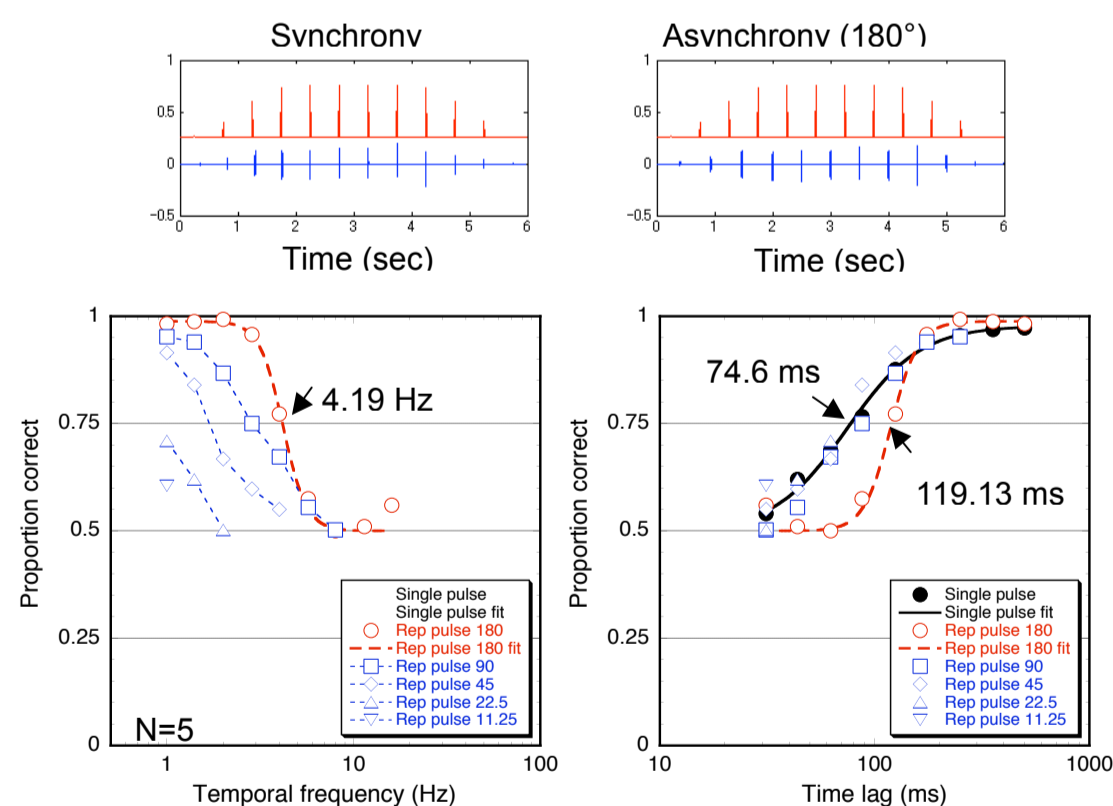


Fig. 2: The proportion correct of audiovisual synchrony-asyncrony discrimination for repetitive pulse trains and for the single pulse condition, plotted as a function of the temporal frequency (left) and the time lags between audio-visual signals (right).

MATCHING FEATURE

Feature tracking motion perception is based on the position change of salient features in the image (Lu & Sperling, 2001, *JOSA*). Similarly, audiovisual synchrony perception is based on the matching of salient temporal features (Fig. 3). These features are extracted from low-level sensory signals either by bottom-up

segregation process, or by top-down attentional process.

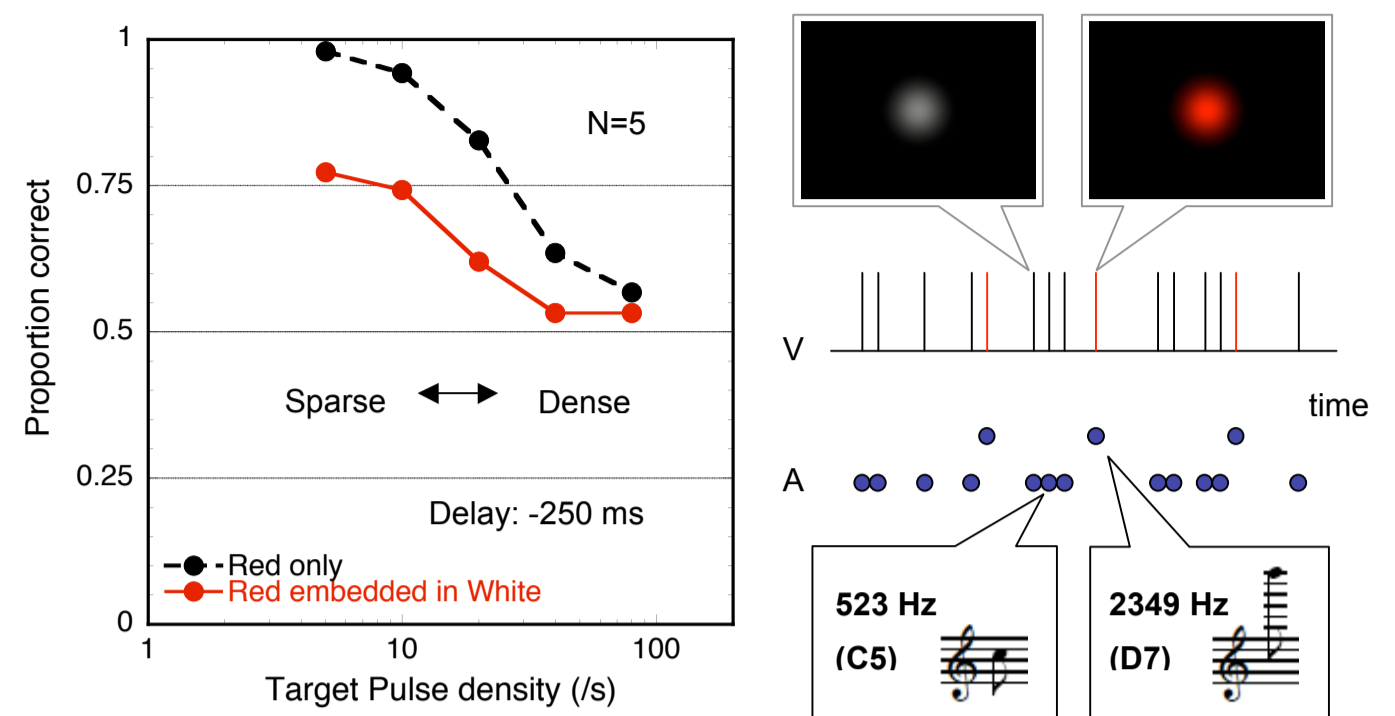


Fig. 3: Audiovisual synchrony-asyncrony discrimination for random pulse trains. Black: The stimulus consisted of figure pulse (red flash and D7 pip) only. Synchrony detection was impossible for the highest pulse density (80 pulse/s). Red: Figure pulse was embedded in background pulse (white flash and C5 pip, right panel). Although the total pulse density was always 80 pulse/s, synchrony detection was possible when the stimulus contained salient features (sparse figure pulse).

VISUAL SEARCH

Visual search for a motion-defined target is serial when the motion is processed by feature tracking mechanism (Ashida *et al.*, 2001, *JOSA*). Similarly, detection of a visual target that changes in synchrony with an auditory stimulus is serial (Fig. 4).

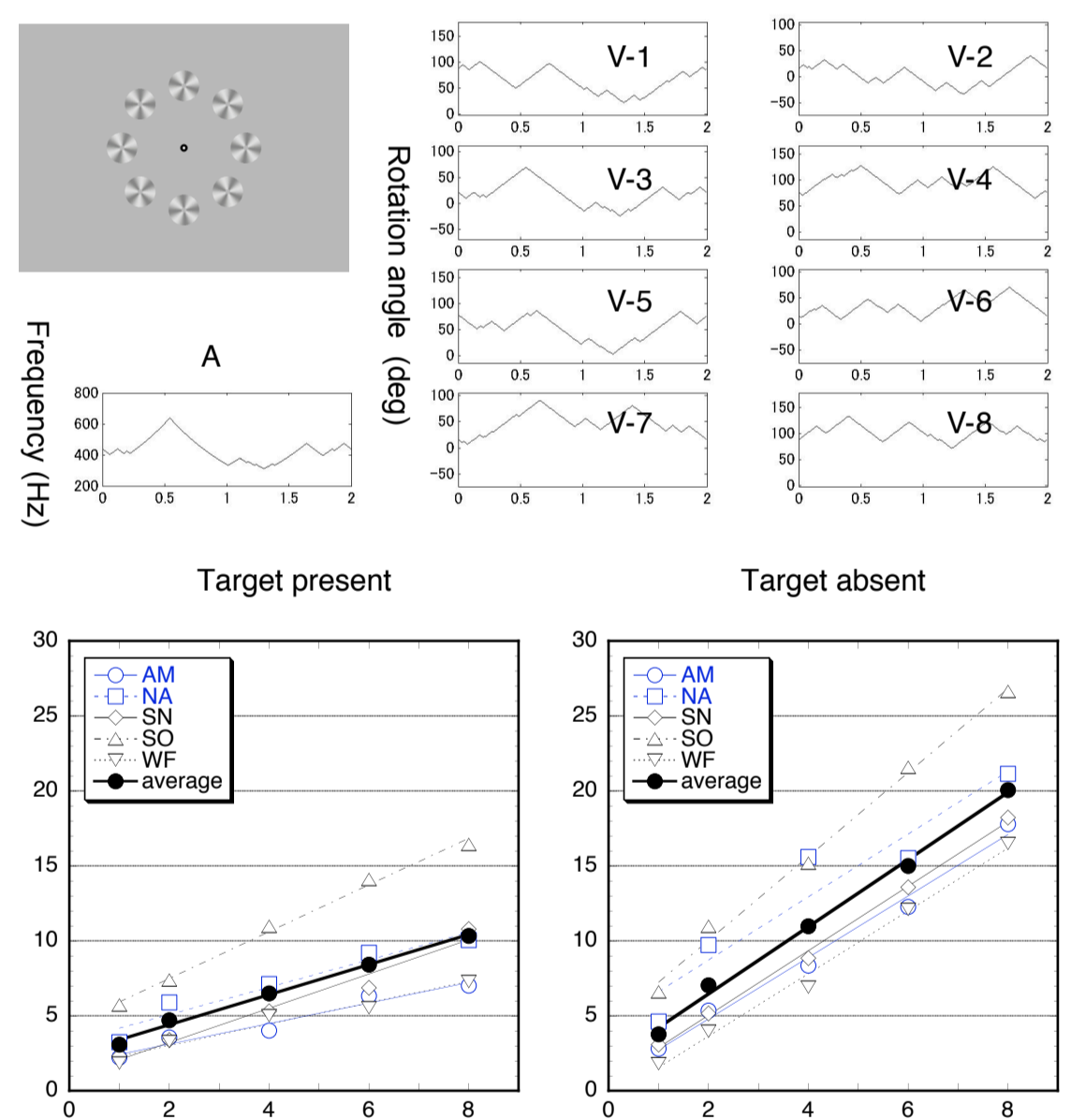


Fig. 4: Visual search for a windmill target whose clockwise-counterclockwise rotation was synchronized with the frequency-modulated up-down sweep of the pure tone auditory signal. Reaction times (RT) for target detection increased linearly with number of distractors, with the slope being about two times as steep for target-absent trials as for target-present trials.

VISUAL TEMPORAL BINDING

Without special neural hardware for feature binding, purely visual synchrony perception is also slow and presumably attentive: colour vs. spatially separate orientation (Holcombe & Cavanagh, 2001, *Nat Neurosci*); colour vs. motion (Nishida & Johnston, 2002, *Curr Biol*); two luminance flickers with a large spatial gap (e.g., Victor & Conte, 2002, *Vision Res*).

CONCLUSION

Salient feature matching may be a common principle of mid-level temporal binding.