Gagnon, L., Foucher, S., Laliberté, F. and G. Boulianne. "A Simplified Audiovisual Fusion Model with application to Large-Vocabulary Recognition of French Canadian Speech = Un modèle simplifié de fusion audio-visuelle avec application à la reconnaissance à grand vocabulaire du français canadien parlé" *Canadian Journal of Electrical and Computer Engineering = Journal canadien de génie électrique et informatique*, 33 (2) Spring/Printemps 2008 : 109-119

Abstract
A new, simple and practical way of fusing audio and visual information to enhance audiovisual automatic speech recognition within the framework of an application of large-vocabulary speech recognition of French Canadian speech is presented, and the experimental methodology is described in detail. The visual information about mouth shape is extracted off-line using a cascade of weak classifiers and a Kalman filter, and is combined with the large-vocabulary speech recognition system of the Centre de Recherche Informatique de Montréal. The visual classification is performed by a pair-wise kernel-based linear discriminant analysis (KLDA) applied on a principal component analysis (PCA) subspace, followed by a binary combination and voting algorithm on 35 French phonetic classes. Three fusion approaches are compared: (1) standard low-level feature-based fusion, (2) decision-based fusion within the framework of the transferable belief model (an interpretation of the Dempster-Shafer evidential theory), and (3) a combination of (1) and (2). For decision-based fusion, the audio information is considered to be a precise Bayesian source, while the visual information is considered an imprecise evidential source. This treatment ensures that the visual information does not significantly degrade the audio information in situations where the audio performs well (e.g., a controlled noise-free environment). Results show significant improvement in the word error rate to a level comparable to that of more sophisticated systems. To the authors' knowledge, this work is the first to address large-vocabulary audiovisual recognition of French Canadian speech and decision-based audiovisual fusion within the transferable belief model.

Keywords : audiovisual speech recognition; Dempster-Shafer theory; mouth tracking; multimodal fusion