MULTIModal SVM for speech recognition

Gabriel Costache, Inge Gavat, Diana Militaru
University Politehnica of Bucharest

Abstract

We present an alternative for building robust speech classification systems, biologically inspired from human perception as a multimodal process, in order to improve the machine understanding in the spoken man-machine interaction. Compared with classical unimodal classification systems this approach uses features obtained from more than one source to characterize the patterns to be classified. The processing of data obtained this way will enhance the possibility of the system to have a robust behavior under degraded input conditions, in order to ensure better performance in the real world, under spoken dialog conditions. Such a classification system will have small variations of classification rates when the features from one source or more sources are corrupted with unwanted components.

Introduction

- Multimodality is biologically inspired. Recognize a pattern using multiple sources of information.
- Automated speech recognition accuracy drops when the speech signal is corrupted with noise.
- Lip-reading is useful by itself to recognize speech and it is more robust to noise comparing with speech signal.
- SVM – statistical binary classification algorithm with good generalisation property, extended to multiclass classification

Multimodal Systems

- Combining multiple sources of algorithms to classify/analyze a pattern
- Feature fusion: combine features before applying classification algorithm
- Decision fusion: combine scores from multiple classification algorithm
- Improve robustness against unimodal features

Support Vector Machines

Binary SVM

- Find optimal hyperplane
- Minimize cost function: \( \Phi(w, \xi) = \frac{1}{2} w^T w + C \sum \xi_i \)

Non linear mapping: Inner product kernel

- Polynomial: \( K(x, x') = \phi(x) \cdot \phi(x') \)
- Radial basis function
- Two layer perceptron

Audio&Video Processing

Speech Processing

Perceptual Linear Prediction

Perceptual Cepstral Analysis

Video Processing

- Face tracking using GMM
- Extract mouth geometric features

Results

Database:
- AMP Carnegie Mellon
- 10 speakers (10 pronunciations)
- Window = 256 samples
- Fs = 8 KHz
- PLP - 16 bank filters
- MFCC - 13 order
- Image: Fps=30
- 3 coefficients: x2-x1, h1, h2
- Artificial noise over speech signal

Conclusion:

- Multimodal features have better results than unimodal ones especially when noise is present over the speech signal.
- Using perceptual features we can obtain better recognition rates compared with the classical features
- MFC coefficients are more robust than PLP ones
- Multiclass SVM has a good generalisation property

References: