Speech Recognition Techniques for a Sign Language Recognition System

Philippe Dreuw, David Rybach, Thomas Deselaers, Morteza Zahedi, and Hermann Ney
Human Language Technology and Pattern Recognition, RWTH Aachen University, Aachen, Germany

Introduction

- automatic sign language recognition system
- necessary for communication between deaf and hearing people
- continuous sign language recognition, several speakers, vision-based approach, no special hardware
- large vocabulary speech recognition (LVSR) system to obtain a textual representation of the signed sentences
- evaluation of speech recognition techniques on publicly available sign language corpus

Automatic Sign Language Recognition (ASLR)

- similar to speech recognition: temporal sequences of images
- important features
  - hand-shapes, facial expressions, lip-patterns
  - orientation and movement of the hands, arms or body
- HMMs are used to compensate time and amplitude variations of the signers
- goal: find the model which best expresses the observation sequence

Experimental Setup

Database

- system evaluation on the RWTH-BOSTON-104 database
  - 201 sentences (161 training and 40 test sequences)
  - vocabulary size of 104 words
  - 3 speakers (2 female, 1 male)
  - corpus is annotated in glosses

Problems

- 26% of the training data are singletons
- simple sentence structure
- one out-of-vocabulary (OOV) words with whole-word models

Differences in Comparison to ASR

- simultaneousness
- signing space
- environment
- speakers and dialects
- coarticulation and movement epenthesis
- silence
- whole-word models and sub-word units

System Overview

Visual Modeling (VM)

- related to the acoustic model in ASR
- HMM based, with separate GMMs, globally pooled diag. covariance matrix
- monophone whole-word models
- pronunciation handling

Language Modeling (LM)

- according to ASR: LM should have a greater weight than the VM
- trigram LM using the SRILM toolkit, with modified Kneser-Ney discounting with interpolation

Features

- appearance-based image features: for baseline system
  - thumbnails of video sequence frames (intensity images scaled to 32x32 pixels)
  - give a global description of all (manual and non-manual) features proposed in linguistic research
- manual features:
  - dominant hand tracking: hand position, hand velocity, and hand trajectory features

Feature Selection and Model Combination

Feature Selection

- concatenation of appearance-based and manual features
- sliding window for context modeling
- dimensionality reduction by PCA and/or LDA

Model Combination

- log-linear combination of independently trained models
- profit from independent alignments (e.g. performing well for long and short words)
- profit from different feature extraction approaches

Experimental Results

Features	Dim.	[%WER]
frame intensity (w/o pronunciations)	1024	54.0
frame intensity (w/ pronunciations)	1024	37.0
frame intensity (w/ pronunciations + tangent distance)	1024	33.7
PCA-frame	110	27.5
PCA-frame, hand-position	112	25.3
PCA-frame, hand-velocity	112	24.2
PCA-frame, hand-trajectory	112	23.6
model-combination	2x100	17.9

Example Results

Correct Examples

IX-1P FIND SOMETHING-ONE BOOK
IX-1P FIND SOMETHING-ONE BOOK

JOHN FISH WONT EAT BUT CAN EAT CHICKEN
JOHN FISH WONT EAT BUT CAN EAT CHICKEN

LOVE JOHN WHO
LOVE JOHN WHO

JOHN BUY YESTERDAY WHAT BOOK
JOHN BUY YESTERDAY WHAT BOOK

Incorrect Examples

MARY VEGETABLE KNOW IX LIKE CORN
MARY VEGETABLE KNOW IX LIKE MARY

JOHN IX GIVE MAN IX NEW COAT
JOHN IX WOMAN ____ __ NEW COAT

LIKE CHOCOLATE WHO
JOHN LIKE CHOCOLATE WHO

JOHN [UNKNOWN] BUY HOUSE
JOHN FUTURE NOT BUY HOUSE

RWTH-BOSTON-104 Database

Corpus Statistics

	Training	Test
sentences	161	40
running words	710	178
frames	12422	3324
vocabulary	103	65
singletons	27	9
OOV	-	1

LM Perplexities

LM type	*PP*
zerogram	106.0
unigram	36.8
bigram	6.7
trigram	4.7

Database is publicly available

Conclusion

- LVSR system is suitable for vision-based continuous sign language recognition
- many of the principles known from ASR can directly be transfered
- important for ASLR: temporal contexts, pronunciation handling, language modelling, and model combination
- outlook: connection of recognizer output to a statistical machine translation system achieved promising translation results