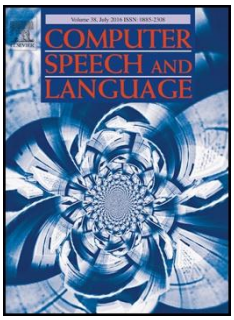
Accepted Manuscript



Title: Cross database audio visual speech adaptation for phonetic spoken term detection

Author: Shahram Kalantari, David Dean, Sridha Sridharan

PII: S0885-2308(15)30013-9
DOI: http://dx.doi.org/doi: 10.1016/j.csl.2016.09.001
Reference: YCSLA 791

To appear in: *Computer Speech and Language*

Received date: 27-9-2015
Revised date: 7-6-2016
Accepted date: 11-9-2016

Please cite this article as: Shahram Kalantari, David Dean, Sridha Sridharan, Cross database audio visual speech adaptation for phonetic spoken term detection, *Computer Speech and Language* (2017), http://dx.doi.org/doi: 10.1016/j.csl.2016.09.001.

Cross database audio visual speech adaptation for phonetic spoken term detection

Shahram Kalantari*, David Dean*, and Sridha Sridharan*

**Speech, Audio, Image and Video Technology Lab, Queensland University of Technology, Australia.*

s1.kalantari@qut.edu.au, ddean@ieee.org, s.sridharan@qut.edu.au

Highlights

- We show that use of visual information helps both phone recognition and spoken term detection accuracy.
- Fused HMM adaptation could be utilised to benefit from multiple databases when training audio visual phone models
- An additional audio adaptation improves cross-database training accuracy for phone recognition and spoken term detection
- A post training step can be used to update all HMM parameters and further improve phone recognition accuracy

Abstract

Spoken term detection (STD), the process of finding all occurrences of a specified search term in a large amount of speech segments, has many applications in multimedia search and retrieval of information. It is known that use of video information in the form of lip movements can improve the performance of STD in the presence of audio noise. However, research in this direction has been hampered by the unavailability of large annotated audio visual databases for development. We propose a novel approach to develop audio visual spoken term detection when only a small (low resource) audio visual database is available for development. First, cross database training is proposed as a novel framework using the fused hidden Markov modelling (HMM) technique, which is used to train an audio model using extensive large and publicly available audio databases; then it is adapted to the visual data of the given audio visual database. This approach is shown to perform better than standard HMM joint-training method and also improves the performance of spoken term detection when used in the indexing stage. In another attempt, the