FIRST EVALUATION OF ACOUSTIC EVENT CLASSIFICATION SYSTEMS IN CHIL PROJECT

Robert Malkin

INTERACT
Carnegie Mellon University
Pittsburgh, USA
Email: malkin@cs.cmu.edu

Dušan Macho, Andrey Temko, Climent Nadeu
TALP Research Center
Universitat Politecnica de Catalunya
Barcelona, SPAIN
Email: dusan@gps.tsc.upc.es

Introduction to the CHIL Project

CHIL is a European Commission sponsored project whose goal is to provide computational systems that deliver services to humans in an unobtrusive fashion. The acronym "Computers in the Human Interaction Loop" reveals the project's focus on adapting machine services to human needs, rather than forcing humans to adapt to the computers that are nominally serving them. See [1] for more details on the CHIL project.

Crucial to the goal of the CHIL project is the ability to determine human context from auditory or visual cues in the environment. Toward this end, several audio and video technologies were developed and evaluated within the project. In this contribution we present systems that identify common acoustic events in the first CHIL scenario, seminars. Systems focusing on other audio aspects of CHIL like speech detection and recognition or speaker localization are presented in [2].

Description of the CHIL Evaluation

To evaluate acoustic event classification (AEC) systems, we used a CHIL seminar database that was collected in 2003 at the University of Karlsruhe and consists of seven technical seminars. Audio was collected with a combination of microphones; we used one channel of a wall-mounted 8-element linear microphone array for AEC evaluation. The data were manually transcribed with 25 noise classes (e.g. keyboard, door, or step noises, but also human noises like cough, throat cleaning, or breathing), and over 2800 individual noise instances were collected. These instances were transcribed with tight temporal bounds, allowing us to perform an isolated-sound test. We plan to move to continuous recognition in mid to late 2005.

The output of the evaluated AEC system was compared with the manually labeled reference. An overall accuracy that measures the percentage of correctly classified events was employed as a metric for the system comparison.

Currently more CHIL seminar data is being collected with more acoustic sensors involved. These new data will be included in the forthcoming second CHIL evaluation round that will be accomplished within the next two to three months. Using the expertise from the first evaluation, the evaluation criteria for AEC were also improved. We plan to present the results from this second round CHIL evaluation at the workshop.

Short Description of AEC Systems

CMU

In the CMU AEC system, we tested several sets of acoustic features with a GMM/HMM classifier. We extracted three baseline sets of acoustic features; logscale mel spectra, cepstra, and a set of perceptual features consisting of spectral brightness, spectral diffusion, and power. The perceptual features were combined with both the log-mel features and the cepstral features. We then combined three-frame and five-frame chunks of these features into temporal feature sets using PCA, ICA, and LDA.

We employed a standard approach of building a simple GMM on all possible features, and then selecting the most promising feature sets with which to build more complex classifiers based on labels from simpler systems. The best GMM systems were based on PCA and ICA feature sets derived from 5 frames of log-mel features, and used acoustic event priors (derived from the training data) weighted at 65%. These systems achieved an accuracy of 58.7%. The final system, a full

GMM/HMM system, used LDA based on 5 frames of log-mel features and acoustic event priors weighted at 10%. This system achieved an accuracy of 61.59%.

More details on these experiments, as well as results on the upcoming CHIL evaluations, will be presented at the workshop.

UPC

The UPC AEC's results presented in this communication have been obtained with two classification systems which are based on either Support Vector Machines (SVM) or Gaussian Mixture Models (GMM). Both classification systems were used in combination with the following acoustic features:

a) A set of perceptual features (zero-crossing rate, pitch, short-term energy, sub-band energies, spectral flux),
b) The log energy feature,
c) Cepstral features and their time derivatives,
d) Frequency-filtered-band-energy features and their time derivatives.

The above feature sets were used both individually and in combinations. After calculating the feature vectors using a frame-by-frame analysis for the entire utterance, the mean, variance, autocorrelation and entropy of all feature vectors were estimated. These four measures formed one vector representing the given utterance (event) and it was used as input to the classifier.

In the first run of CHIL evaluations, the best AEC accuracy for the UPC GMM system was obtained by using the Mel-cepstrum feature set in combination with the log energy feature (47.4%). As for the UPC SVM system, a combination of the perceptual features with the frequency filtering features obtained the best performance (55.1%).

At the time of writing of this abstract, we have improved versions of the AEC classification systems which we will test on the new CHIL evaluation data. They use a scheme with a specific feature set at each node. For training, a tree clustering technique is employed that relies on a given set of confusion matrices and chooses the most discriminative feature set at each step of classification (see [3] for more details). The new results will be presented at the workshop as well.

References

[1] A. Waibel et al., "CHIL: Computers in the Human Interaction Loop", in International Workshop on Image Analysis for Multimedia Interactive Services (IWIAMIS 2004), 2004.

[2] D. Macho et al., "First experiments of automatic speech activity detection, source localization and speech recognition in the CHIL project," submitted to HSCMA 2005.

[3] A. Temko, C. Nadeu, "Meeting room acoustic event classification by support vector machines and variable-feature-set clustering", accepted to *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, 2005