AUDIO-BASED CONTEXT AWARENESS – ACOUSTIC MODELING AND PERCEPTUAL EVALUATION

Antti Eronen[*], *Juha Tuomi*[*], *Anssi Klapuri*[*], *Seppo Fagerlund*[+], *Timo Sorsa*[+], *Gaëtan Lorho*[+], *Jyri Huopaniemi*[+]

[*]Tampere University of Technology
Institute of Signal Processing
P.O.Box 553, FIN-33101 Tampere, Finland
`antti.eronen@tut.fi`

[+]Nokia Research Center
Speech and Audio Systems Laboratory
P.O.Box 407, FIN-00045 Nokia Group,Finland
`timo.sorsa@nokia.com`

ABSTRACT

This paper concerns the development of a system for the recognition of a context or an environment based on acoustic information only. Our system uses mel-frequency cepstral coefficients and their derivatives as features, and continuous density hidden Markov models (HMM) as acoustic models. We evaluate different model topologies and training methods for HMMs and show that discriminative training can yield a 10% reduction in error rate compared to maximum-likelihood training. A listening test is made to study the human accuracy in the task and to obtain a baseline for the assessment of the performance of the system. Direct comparison to human performance indicates that the system performs somewhat worse than human subjects do in the recognition of 18 everyday contexts and almost comparably in recognizing six higher level categories.

1. INTRODUCTION

Information about the environment would enable wearable devices to provide better service to users' needs, e.g. by adjusting the mode of operation according to the context. Many sources of information for sensing the environment are available. In this paper, we consider *audio-based context awareness*, where the decision is based merely on the available acoustic information.

The reported work continues the research first described in [1]. In this paper, we describe a listening test made to facilitate the direct comparison of the system's performance to that of human subjects. A forced choice test with identical test samples and reference classes for the subjects and the system is used. The second main concern in this paper is to evaluate different methods for training the hidden Markov models used to represent the feature statistics. We are dealing with a highly varying acoustic material where practically any imaginable sounds can occur. Thus, it is most likely that the acoustic models we are using are not able to sufficiently model the observation statistics. Therefore, we propose using discriminative training instead of conventional maximum-likelihood training. Neither is there any guarantee that a similar model would be appropriate for all classes. Determining the structure of the model from the data is thus an interesting approach, and we make experiments with an algorithm recently proposed for simultaneous training and model order selection of one-state HMMs [2].

2. ACOUSTIC MEASUREMENTS

The database consisted of 225 real-world recordings from a variety of different contexts, or environments. The recording procedure has been described in [1]. Two training and testing setups were formed from the samples. Setup 1 consisted of 155 recordings of 24 contexts that were used for training and 70 recordings of 16 contexts were tested. Setup 2 was used in the listening test and in the direct comparison, and had two non-overlapping sets of 45 samples from 18 different contexts in the test set. A higher level of abstraction may be sufficient for some applications. Hence, the recordings were also categorized into six classes that are more general according to some common characteristics. These categories are outdoors, vehicles, public places, offices and quiet places, home, and reverberant places.

3. SYSTEM DESCRIPTION

3.1. Feature extraction

Mel-frequency cepstral coefficients (MFCC) were found to be a well performing feature set in this task [1], and are used as the front-end parameters in our system. The input signal is first pre-emphasized with the FIR filter $1, -0.97z^{-1}$. MFCC analysis is performed in 30 ms windowed frames advanced every 15 ms. The number of triangular filters was 40, and they occupied the band from 80Hz to half the sampling rate. The number of cepstral coefficients was 11 after the zeroth coefficient was discarded, and appending the first time derivatives approximated with a 3-point first-order polynomial fit resulted in a feature vector size of 22. The resulting features were both mean and variance normalized.

3.2. The hidden Markov model

A continuous density hidden Markov model (HMM) with N states consists of a set of parameters θ that comprises the N-by-N transition matrix, the initial state distribution, and the weights, means and diagonal variances of Gaussian mixture model (GMM) state emission densities. Training is performed in the training set that consists of the recordings $\mathrm{O} = (\mathbf{O}^1, ..., \mathbf{O}^R)$ and their associated class (context) labels $L = (l^1, ..., l^R)$. Specifically, $\mathbf{O}^r$ denotes the sequence of feature vectors measured from recording r. Typically, the HMM

parameters are iteratively optimized with the Baum-Welch algorithm that finds a local maximum of the maximum likelihood (ML) objective function

$$F(\Theta) = \sum_{c=1}^{C} \sum_{r \in A_c} \log p(\mathbf{O}^r \mid c),$$

where Θ denotes the entire parameter set of all the contexts $c \in \{1,...,C\}$, and A_c is the subset of $[1,R]$ that denotes the recordings from the context c. In the recognition phase, an unknown recording $\mathbf{O}$ is classified using the maximum *a posteriori* rule:

$$\hat{c} = \arg\max_c p(\mathbf{O}^r \mid c).$$

The needed likelihoods can be efficiently computed using the forward-backward algorithm, or approximated with the likelihood of the single most likely path given by the Viterbi-algorithm.

3.3. Discriminative training

ML estimation is well justified if the observations are distributed according to the model. If a model mismatch occurs, other approaches may lead into better performing models. Discriminative training methods such as the maximum mutual information (MMI) aim at maximizing the ability to distinguish between the observation sequences generated by the model of the correct class and those generated by models of other classes [5]. The MMI objective function is given as

$$M(\Theta) = \log p(L \mid O) = \sum_{r=1}^{R} \log p(l^r \mid \mathbf{O}^r)$$
$$= \sum_{r=1}^{R} \left\{ \log\left[p(l^r) p(\mathbf{O}^r \mid l^r) \right] - \log \sum_{c=1}^{C} p(c) p(\mathbf{O}^r \mid c) \right\},$$

where $p(l^r)$ and $p(c)$ are prior probabilities. Unfortunately, there exists no simple optimization method for this problem. The optimization involves the entire model set even if only observations from a single class were used.

Different discriminative algorithms have been proposed. The one used in this paper is perhaps the most straightforward to implement. The algorithm was proposed by Ben-Yishai & Burshtein, and is based on the *approximated maximum mutual information* (AMMI) criterion [6]. Their criterion is:

$$J(\Theta) = \sum_{c=1}^{C} \left\{ \sum_{r \in A_c} \log\left[p(c) p(\mathbf{O}^r \mid c) \right] - \lambda \sum_{r \in B_c} \log\left[p(c) p(\mathbf{O}^r \mid c) \right] \right\},$$

where B_c is the set of indices of training recordings that were *recognized* as c. B_c is obtained by maximum a posteriori classification performed on the training set. The parameter $0 \le \lambda \le 1$ controls the "discrimination rate".

The prior probabilities $p(c)$ do not affect the maximization of $J(\Theta)$, thus the maximization is equivalent to maximizing the following objective functions:

$$J_c(\Theta) = \sum_{r \in A_c} \log p(\mathbf{O}^r \mid c) - \lambda \sum_{r \in B_c} \log p(\mathbf{O}^r \mid c),$$

for all $1 \le c \le C$. Thus, the parameter set of each context can be estimated separately, which leads to a straightforward implementation. The authors give the re-estimation equations for HMM parameters [6]. Due to space restrictions, we present only the re-estimation equation for the transition probability from state i to state j:

$$\bar{a}_{ij} = \frac{\sum_{r \in A_c} \sum_{t=1}^{T_r-1} \xi_t(i,j) - \lambda \sum_{r \in B_c} \sum_{t=1}^{T_r-1} \xi_t(i,j)}{\sum_{r \in A_c} \sum_{t=1}^{T_r-1} \gamma_t(i) - \lambda \sum_{r \in B_c} \sum_{t=1}^{T_r-1} \gamma_t(i)},$$

where $\xi_t(i,j) = p(q_t = i, q_{t+1} = j \mid \mathbf{O}^r, c)$ and $\gamma_t = \sum_{j=1}^{N} \xi_t(i,j)$. The state at time t is denoted by q_t, and the length of the observation sequence $\mathbf{O}^r$ is T_r. In a general form, for each parameter v the re-estimation procedure is

$$v = \frac{N(v) - \lambda N_D(v)}{D(v) - \lambda D_D(v)}$$

where $N(v)$ and $D(v)$ are accumulators that are computed according to the set A_c, and $N_D(v)$ and $D_D(v)$ are the discriminative accumulators computed according to the set B_c, obtained by recognition on the training set. This discriminative re-estimation can be iterated, we used typically 5 iterations.

3.4 Simultaneous training and order selection of GMMs

A one-state continuous-density HMM, i.e. a GMM, is treated here separately since recently an interesting algorithm for simultaneous training and order selection of GMMs has been presented [2]. The algorithm is based on embedding the evaluation of an information theoretic criterion, in this paper the Minimum Description Length (MDL) within the EM-iterations. The criterion includes a term measuring the likelihood of the data, and a penalizing term that grows, as the models become more complex. The algorithm starts with a large number of components $M_{max} \gg 1$, and then merges or destroys components with little support, evaluating the criteria for each model candidate. The final model is the one giving the minimum for the criterion. This algorithm is referred as the *agglomerative EM* (AEM).

In our implementation of the algorithm, the search for the components to be merged was slightly modified from the procedure presented in [2]. After the EM converged, the component m_{min} with the smallest weight was merged into the component that corresponded to the minimum of the symmetric Kullback-Leibler divergence between m_{min} and the other densities.

Table 1. Recognition rates with Setup 1 using GMMs.

# of components	Baseline	AMMI
$M = 1$	66.4	68.3
$M = 2$	70.6	73.2
$M = 3$	67.4	67.4

Table 2. Recognition rates with Setup 1 using HMMs.

	Fully connected		Left-right	
States	ML	AMMI	ML	AMMI
2	74.7	77.3	-	-
3	73.7	73.7	74.5	75.0
4	68.0	71.9	71.1	72.7

4. RESULTS

4.1. Test setup

Two test setups were used. In Setup 1, 70 samples of 16 different contexts were tested, and models for 24 reference classes were trained using all the 155 samples not included in the test set. In the training set, the length of samples was 160 s, in the test set only 60 s was used in all simulations.

For Setup 2, models were trained separately for both 45-sample subsets, and each time all the remaining samples were used for model training. The final recognition result was the average of the two results.

4.2. GMM

The results obtained with GMMs trained with different algorithms are presented in Table 1. The baseline GMM recognition system consisted of a fixed-order model trained with the EM-algorithm. We trained models with different number of components (M). The best accuracy was obtained with just two Gaussian components. Table 1 also shows the results obtained using the discriminative training algorithm (AMMI). Values of λ ranging from 0.1 to 0.9 were tested, and the number of discriminative training iterations was 5. In Table 1, the recognition rate obtained by using the model corresponding to the *best training set recognition accuracy* is reported. The optimal value for λ was between 0.3 and 0.4, and the number of discriminative training iterations giving the best training set accuracy varied between 2 and 5. It was observed that usually the maximal test set accuracy was obtained with value of λ and number of iterations other than those resulting in the best accuracy on the training set. However, since we do not have access to the test set in real applications, those can only be considered as the upper limit of achievable recognition rate, and are 71.9%, 73.7% and 72.1% for 1, 2, and 3 components, respectively.

The size of our database compared to the dimension 22 of the feature vector proved out to be too small for the AEM algorithm to work properly: it gave considerably larger model orders than the optimal baseline with $M = 2$. We wanted to test if AEM could provide any performance gain if lower-dimensional features were used. We downsampled the data to 8 kHz, and used only 7 dimensional MFCC coefficients as features. Now, the AEM gave an average model order of 38.6 with standard deviation 3.98. The baseline with 20 components provided an accuracy of 64.1%, whereas the models trained with the AEM resulted in 62.6% accuracy.

4.3 HMM

The Baum-Welch algorithm was used to train the baseline HMMs. The state means and variances were initialized by *k*-means clustering. The topologies tested were a fully connected HMM and a left-right HMM with skips. The number of states and component densities per state was varied. Increasing the number of components in each state was obtained by gradually increasing the model order from one to the desired order by splitting the component with the largest weight. The results in Table 2 are shown for one-component emission densities that yielded the best recognition accuracy. Increasing the number of components in state emission densities lowered the recognition rates somewhat.

Table 2 also shows the results obtained using discriminative training (AMMI). As with GMMs, different values of λ and different number of discriminative training iterations were tested, and Table 2 shows the recognition rate obtained with models giving the best accuracy on the training set.
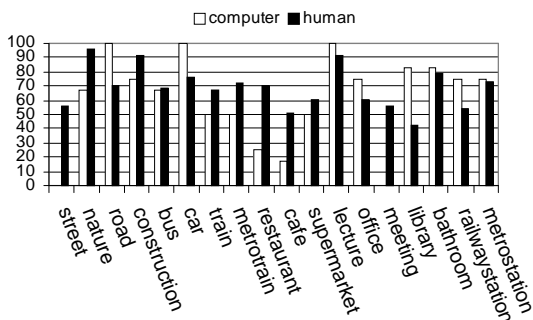
5. LISTENING TEST

The aim of the listening test was to study the human ability to recognize contexts based on auditory signals only, in order to obtain a baseline for the assessment of computational model performance. This experiment was organized in three listening tests. The first studied how accurately and how fast the environment could be recognized. The second considered the importance of spatial information in recognition. The third investigated the importance of different cues for recognition.

5.1 Listening test setup

All tests were performed in an ITU-R BS.1116-1 [3] compliant listening room. Audio samples were reproduced over a stereophonic setup using Genelec 1031A loudspeakers placed at ±30° in front of the listener. The test design and administration were performed using the Presentation software [4].

18 subjects participated in the test, which was designed for two groups, each including the same number of stimuli and identical contexts. This permitted the use of a larger amount of samples, while keeping the total duration of the test within one hour per subject. The listening test started with a training session using nine samples not included in the actual test to familiarize the subjects with the user interface and the test setup. In the three experiments, subjects were instructed to try to recognize the context as fast as possible. The subjects were asked to make a forced choice from the list of 27 possible responses.

Figure 1. *Comparison of recognition rates for the 18 contexts.*

5.2 Results of the listening test

Two measures were analyzed from this test, the recognition rate and the reaction time for each stimulus. The recognition rate was analyzed as a set of right or wrong answers using a non-parametric statistical procedure. For the reaction time, the statistical analysis was performed with a parametric statistical procedure (ANOVA).

5.2.1 Stereo test
The average recognition rate was 69% for context and 88% allowing confusions within the six higher level categories. Figure 1 presents the recognition rate for each of the 18 contexts averaged over all listeners (differences between groups were not significant). Overall, the average reaction time was 13 seconds ranging from 5 seconds (nature) to 21 seconds (library). It should be noted, however, that reaction time for the higher level recognition only would probably be significantly faster. Indeed, some of the subjects reported that they could exclude most of the contexts fast, but the final decision between specific contexts from the same broader class took more time.

5.2.2 Mono/Stereo/Binaural test
For this test, recognition rates were compared for monophonic, stereophonic and binaural presentations. A set of 18 samples not included in the stereo test from nine different environments was used for each configuration. For the binaural samples, cross-talk cancellation was applied. The recognition rate averaged over the three techniques was 66% for context and it increased to 88% for the higher level categories. The average recognition rate for binaural, mono and stereo samples were 62%, 63% and 70% at the context level, respectively. For the higher level categories, the rates were 90%, 86% and 89%. These differences are not statistically significant. However, differences in reaction times were significant.

5.2.3 Qualitative test
In the last sub-test, subjects were asked to listen to nine samples and rate the information they used in the recognition process. After each stimulus, listeners filed a form in which they were asked to evaluate and rate on a 6-point discrete scale the importance of different cues in recognition. The results indicate that human activity and spatial information cues are most often used. Two cues that were not so often used but received high importance ratings were prominent events and nature sounds.
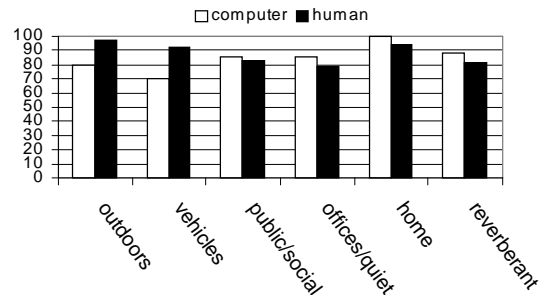
Figure 2. *Comparison of recognition rates on the six higher level categories.*

6. COMPARISON TO THE HUMAN ABILITY

A direct comparison to the human ability was made using exactly the same test samples and reference classes as in the listening test. Figures 1 and 2 summarize the results for the subjects and the system on this test setup. Results are shown both for context recognition and at the higher category level. The results of the system have been obtained using 2-state fully-connected HMMs. The averaged recognition accuracies of the computer system are 61% and 85% against the accuracies 69% and 88% obtained in the listening test for context and higher level classes, respectively. No improvement was observed by using discriminative training; the results with models giving the best training set accuracy were the same as with the baseline. However, the upper-limit recognition rates were 64% and 90% for 18 and 6 classes, respectively, with $\lambda = 0.4$ and four iterations of discriminative training. One explanation for the differences in results at the context level may be that the system does not use any spatial information.

REFERENCES

[1] V. Peltonen, J. Tuomi, A. Klapuri, J. Huopaniemi, T. Sorsa. "Computational auditory scene recognition". In Proc. IEEE ICASSP 2001.
[2] M.A.T. Figueiredo, M. N. Leitao, A.K. Jain. "On Fitting Mixture Models". In Energy Minimization Methods in Computer Vision and Pattern Recognition, E. Hancock and M. Pellilo (Eds.), pp. 54-69, Springer-Verlag, 1999.
[3] ITU-R, Recommendation BS.1116-1. "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems". ITU-R 1997.
[4] Neurobehavioral Systems, "Presentation" [Software].
[5] L. R. Rabiner, B.-H. Juang. *Fundamentals of Speech Recognitio*n, PTR Prentice-Hall Inc., New Jersey, 1993.
[6] A. Ben-Yishai, D. Burshtein, "A Discriminative Training Algorithm for Hidden Markov Models". Submitted to *IEEE Transactions on Speech and Audio Processing*.