Temporal ICA for Classification of Acoustic Events in a Kitchen Environment

Florian Kraft, Robert Malkin, Thomas Schaaf, Alex Waibel

interACT: Universität Karlsruhe; Carnegie Mellon University
Karlsruhe, Germany; Pittsburgh, USA
fkraft@ira.uka.de, {malkin, tschaaf, ahw}@cs.cmu.edu

Abstract

We describe a feature extraction method for general audio modeling using a temporal extension of Independent Component Analysis (ICA) and demonstrate its utility in the context of a sound classification task in a kitchen environment. Our approach accounts for temporal dependencies over multiple analysis frames much like the standard audio modeling technique of adding first and second temporal derivatives to the feature set. Using a real-world dataset of kitchen sounds, we show that our approach outperforms a canonical version of this standard front end, the mel-frequency cepstral coefficients (MFCCs), which has found successful application in automatic speech recognition tasks.

1. Introduction

Recognizing acoustic events is becoming a key component of multimedia computational systems of all types, including personal diaries [1], [2] context-aware mobile devices [3], content-based information retrieval systems, and humanoid robots. An example of such a system is the humanoid robot being developed as part of the SFB 588 project on humanoid robots [4]. This robot is intended to assist elderly or disabled humans in kitchen tasks such as cooking and cleaning, and to provide safety assurance. The ability to detect important kitchen sounds is vital to this set of functions; many important state indicators in the kitchen, like alarms, bells, buzzers, water boiling, or oil beginning to sizzle in a pan, leave little or no visual evidence.

Toward the goal of providing a humanoid robot capable of functioning in a kitchen environment, we have developed a kitchen sound recognition system using a novel feature extraction method based on Independent Component Analysis (ICA). Our method learns ICA basis functions over a multi-frame window of features; these functions capture inter-frame temporal dependencies in an efficient manner. We have evaluated this approach with a corpus of sounds which we collected in real kitchens. We trained standard GMM/HMM classifiers using both our feature extraction approach and a more standard front end which uses Mel-Frequency Cepstral Coefficients (MFCCs) plus first and second temporal derivatives. Our experiments showed that in all cases temporal ICA is a better feature set for describing real-world sounds of this type.

The remainder of this paper is organized as follows. We discuss related work in Section 2, followed by the details of our feature extraction method in Section 3. We then review the data collection procedures in Section 4 before describing the experimental results in Section 5. We conclude with discussion and future directions in Section 6.

2. Related Work

The body of work on identifying real-world acoustic events (as opposed to the related but distinct problem of identifying acoustic events in heavily-produced multimedia corpora like films or broadcast news) is relatively small, but ever-growing, as researchers continue to find both novel applications and methods. [5] gave a good overview of the general audio signal classification problem. [1] detailed an audiovisual processing system capable of capturing and learning patterns in a user's daily life. [6] presented a layered HMM system using PCA features which could detect certain classes of human activity in an office setting. [7] discussed the use of discriminatively-trained HMMs to model general sounds. More recently, [8] and [9] explored various feature sets and classifiers in the context of a meeting room sound recognition system. [10] also explored various types of feature extraction techniques for general audio, with the goal of segmenting long-term recordings into homogeneous chunks. An interesting new feature representation, applicable to short-term sound recognition and potentially long-term environment modeling called the audio Epitome, was presented in [11].

In the speech recognition community, it has been demonstrated that feature sets which model inter-frame dependencies can outperform feature sets which lack this information. Typically, these dependencies have been captured by using the first and second temporal derivatives of the baseline features; Yu [12] showed that performing a Linear Discriminant Analysis (LDA) [13] on sequences of frames (frame-stacking) is superior to this method.

There are many researchers working on ICA; some have used time-domain ICA to learn optimal bases for natural sounds ([14] [15]), while others ([16]) have proposed that ICA be used as an organizing principle for research on computational audition.

3. Feature Extraction using Temporal ICA

Per the result from Yu [12], we wished to capture dependencies between features at timescales longer than one frame. Yu's method used sliding windows (or equivalently, multi-frame stacks) of features as input to LDA. LDA attempts to find a linear transformation of the stacked features such that in the transformed feature space, some set of classes is maximally separable. Under certain conditions [13], LDA can be viewed as an optimal transformation for a classification task. However, LDA has two features which made it ill-suited for our task. First, accurate data segmentation is required at the HMM state level. For new tasks, this kind of segmentation is often not available without a significant amount of effort from a human expert listener. Second, the LDA solution is specific to a set of classes and model topologies, meaning that it is difficult to compare re-

sults across different topology choices and difficult to change the system vocabulary.

Given these problems, we chose to use ICA as an alternative optimal transformation. ICA requires no labels and, as shown by [15], is cabaple of learning a set of functions which correspond quite strongly to actual environmental events.

In the general case, ICA is a family of methods which seek linear transformations of the input data such that the output features are maximally mutually independent [17]. The ICA problem is usually formulated as in Equation 1. Here, some set of underlying, independent signals s is transformed by some unknown linear mixing process A, resulting in the set of observable signals x. Given x, the goal of ICA is to recover both A and s.

$$x = As \tag{1}$$

For the task described here, the vectors x are stacks of observable feature vectors. The columns of A represent features in the transformed space, and the entries s_i represent feature coefficients to be used in our classification system. Hence, in our system, the features s are calculated as $s = xA^{-1}$.

We used Hyvärinen's fastICA [18] procedure to compute the ICA solution A. Though fastICA does not require it, we prewhitened the input data by first using Principal Component Analysis (PCA) [13] to decorrelate the data, followed by a multiplication by the inverse of the square root of the eigenvalue matrix (resulting in an identity covariance matrix). At the same time, we reduced the dimensionality of the feature space by discarding features with low eigenvalues.

In our experiments, we computed ICA solutions for one-frame, three-frame, five-frame, seven-frame, and nine-frame stacks of log melscale spectra. Sets of basis functions for one-frame ICA and seven-frame ICA are shown in Figure 1 and Figure 2. In these figures, the vertical axis corresponds to mel-frequency bins, while within each seven-frame basis, the horizontal axis corresponds to time. Note in Figure 2 that several of the seven-frame bases exhibit strong temporal patterns, and only a few appear to be completely static.

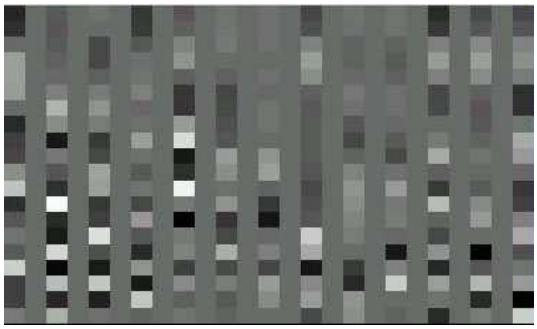

Figure 1: Single-Frame ICA Basis Functions

4. Data Collection and Labeling

To evaluate our system, we collected audio data in four different kitchens. We used a Sony ECM-719 stereo microphone and a Sony MZ-NH700 High-Minidisc recorder. The data were recorded at 44.1KHz and transferred to computer in a lossless fashion. We collected roughly 6000 instances of various kitchen sounds and labeled them by hand. These instances were divided at random into training and test sets; we used 70% of the instances for training and the remainder for testing. We initially divided the sounds into 56 different classes; however, we found that this class set was simply too detailed. Many sets of sound classes were both difficult to distinguish acoustically and lacked semantic import (e.g. "pan on ceramic stove" vs. "pan on metal stove"). After merging these classes, we were left with a 21-class dataset. The dataset is described in Table 1. Here, "telephone" refers to a number of different land and mobile devices, "stove_error" refers to an alarm sound made by a stove, "boiling" refers to simmering water, and "overboiling" refers to water which has reached a rolling boil. The "others" class encompasses silence, stationary background noise, and short, unidentifiable transient noises.

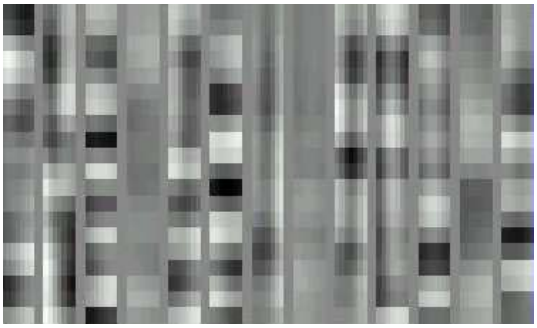

Figure 2: Seven-Frame ICA Basis Functions

Table 1: *Sample counts and durations per class.*

class	# training ex. (dur. in sec)	# test ex. (dur. in sec)	total # ex. (dur. in sec)
boiling	221 (662)	98 (319)	319 (981)
bread_cutter	25 (40)	11 (27)	36 (67)
cutting_vegetables	134 (89)	58 (41)	192 (130)
door	114 (101)	50 (44)	164 (144)
door_bell	50 (110)	22 (55)	72 (164)
egg_timer_ring	11 (34)	6 (17)	17 (51)
footsteps	240 (140)	104 (66)	344 (206)
lighter	84 (42)	37 (20)	121 (61)
match	141 (131)	62 (59)	203 (189)
microwave_beep	110 (30)	49 (17)	159 (47)
others	858 (1130)	369 (547)	1277 (1677)
oven_switch	472 (133)	208 (60)	680 (194)
oven_timer	12 (16)	6 (8)	18 (24)
overboiling	186 (129)	81 (70)	267 (199)
pan_stove	584 (308)	256 (132)	840 (439)
pan_sizzling	107 (343)	46 (146)	153 (489)
telephone	134 (920)	63 (393)	197 (1313)
speech	125 (82)	55 (38)	180 (120)
stove_error	18 (12)	8 (5)	26 (17)
toaster	119 (92)	53 (46)	172 (138)
water	421 (1129)	184 (464)	605 (1593)
total	4166 (5670)	1826 (2573)	5992 (8243)

5. Experiments

To evaluate the efficacy of the temporal ICA features, we performed several classification experiments on the kitchen data. We describe these here.

5.1. Feature Extraction

Using the 44.1KHz audio signal, we first computed a Short-Time Fourier Transform (STFT) with 20ms windows overlapping by 10ms. For the baseline system, BASE, we derived from the short-term power spectra 40 log mel spectra. We then applied the discrete cosine transform (DCT), resulting in 13 MFCCs. We added the first and second temporal derivatives, resulting in a 39-dimensional feature space. For the test systems, we derived 20 log mel spectra from the power spectra and processed them using the temporal ICA procedures described in Section 3 above. For all ICA window lengths, we used a 13-dimensional transformation. In all cases, this allowed us to retain at least 95% of the total eigenvalue mass as calculated via PCA.

5.2. Model Selection and Training

We evaluated several models using the features described in Section 5.1. First, we tested diagonal-covariance Gaussian Mixture Models (GMMs), which contain no temporal structure. Second, we tested Hidden Markov Models (HMMs) with gaussian mixture states. We evaluated both ergodic (ERG) and forward (FWD) HMM topologies with two, three, and four states. To simplify the experiment, each system used the same topology for every class.

In both the GMM and HMM cases, we conducted two separate experiments. In the first, we used the Bayesian Information Criterion (BIC) to choose the optimal number of gaussians per mixture or state. In the second, we kept the number of gaussians per class fixed, dividing the components among states evenly in the HMM case. The average number of gaussians per class for the BIC condition (GMM and ERG models only) is shown in Table 2. Note that the baseline feature set, BASE, has three times as many features as the ICA feature sets. Hence, an ICA system with three times as many gaussians as an BASE system actually has the same number of parameters. After selecting the model size, we initialized the gaussians using the k-means algorithm and then performed Viterbi training [19] to fit gaussian parameters. We did not train the transition probabilities. We performed Viterbi training with optional silence for GMMs as well as HMMs, in order to ameliorate the effects of possibly errorful human segmentations.

Table 2: *Average number of gaussians per class, by system*

Topo	BASE	ICA1	ICA3	ICA5	ICA7	ICA9
GMM	2.5	9.0	10.0	10.0	15.1	9.7
ERG-2	3.0	12.0	11.4	14.2	15.0	13.6
ERG-3	4.5	12.9	14.1	16.5	15.0	15.6
ERG-4	5.2	14.4	16.0	17.2	18.0	18.0

5.3. Model Evaluation and Results

After training, we evaluated the models using the maximum likelihood criterion; each model was exposed to each test example, and the model producing the best likelihood was chosen as the hypothesis. We computed for each system both the average per-class error and the average per-class precision. Results for GMM and HMM systems using BIC-derived optimal gaussian mixture sizes are discussed first, followed by results for systems using fixed gaussian mixture sizes.

5.3.1. GMM Results

GMM results are shown in Table 3. We show unweighted per-class averages for both precision and error. The best performance in terms of both error and precision is achieved by the ICA7 feature set. Note that all temporal ICA feature sets outperform the baseline in terms of error. The ICA7 GMM has twice the number of parameters as the baseline, which may help to explain the 7% difference in error; ICA9, however, only has 29% more parameters than the baseline and a 6% difference in error.

Table 3: *GMM Evaluation results*

Feature Set	Errorl	Precision
BASE	17.3%	80.6%
ICA1	13.2%	79.7%
ICA3	11.8%	80.7%
ICA5	11.2%	80.2%
ICA7	**10.8%**	**83.4%**
ICA9	11.8%	82.3%

5.3.2. HMM Results

HMM results for ergodic models only are shown in Table 4 and 5. Again we see that the best results are achieved by using the ICA7 feature set, with an overall best performance of 9.4% error and 85.5% precision.

As in the GMM condition, some of the variation in performance can be explained by differing mixture sizes. However, the differences in parameter in the HMM case are not as extreme as in the gaussian case. For the 3-state ergodic model, the BASE feature set uses 4.5 gaussians, or 351 parameters per class. The ICA7 feature set uses 15 gaussians, or 390 parameters per class. This is a difference of 11% in terms of parameters; the error difference is 5% in favor of ICA7.

Not shown in these tables are results using forward HMM topologies. Using forward topologies for all classes resulted in worse performace than using ergodic topologies for all classes. Given the makeup of our database, with many sounds having non-forward temporal structure, this is unsurprising. In the future, we intend to report results using optimal topologies per class.

Table 4: *HMM error by topology and feature set*

System	2-state	3-state	4-state
BASE	20.2%	14.4%	16.1%
ICA1	12.2%	11.5%	11.5%
ICA3	11.9%	10.8%	11.7%
ICA5	11.8%	11.5%	11.1%
ICA7	**10.7%**	**9.4%**	**10.5%**
ICA9	11.8%	10.4%	11.3%

5.3.3. Fixed-Gaussian Experiments

For this experiment, we kept the number of gaussians per class fixed at 15. For the 3-state HMM, the gaussians were distributed evenly among the states. Note that this actually means that there are three times as many parameters in the BASE systems as for the ICA systems. Here again, the ICA systems outperform the baseline. Note, however, that the BASE systems now perform

Table 5: *HMM precision by topology and feature set*

System	2-state	3-state	4-state
BASE	79.3%	83.5%	82.2%
ICA1	81.2%	82.5%	82.2%
ICA3	81.8%	82.1%	82.6%
ICA5	83.1%	81.7%	84.4%
ICA7	**84.0%**	**85.5%**	**84.8%**
ICA9	83.5%	84.7%	85.7%

much more reasonably compared to ICA. Note also that the performance difference between GMM and HMM has essentially evaporated; ICA7 even has a test set error 1% *better* for the GMM classifier than the 3-state HMM.

Table 6: *Error and precision, number of gaussians fixed at 15 per class*

System	GMM		ERG3	
	Error	Precision	Error	Precision
BASE	12.4%	80.6%	12.2%	82.8%
ICA1	10.6%	82.8%	10.9%	82.2%
ICA7	**9.2%**	**85.0%**	**10.2%**	**83.4%**

6. Discussion and Future Work

We have presented a sound event classification system which uses ICA to capture inter-frame temporal dependencies in a frame-based feature set. We have demonstrated via experiments using real-world sounds collected in a kitchen setting that this feature extraction method results in better classifier performance than the standard MFCC with first and second temporal derivatives. We have found empirically that ICA over seven frames, or 80ms, results in the best performance for this task.

Our use of HMMs as opposed to other types of classifiers, means that the system is amenable to online, realtime use. Deploying the system in a robot mockup for experimentation with real users in a real kitchen performing real tasks is the next step in evaluating this technology.

Several areas were left unexplored in this paper, but should be revisited in due course. First is some means of automatic topology selection, as suggested in [7]. Second is the use of perceptually-motivated features as input to the temporal ICA procedure, as suggested in [9]. Finally, higher-order knowledge about the nature of human activity scenarios in the kitchen might be exploited in order to provide realistic prior distributions over sequences of acoustic events.

7. Acknowledgments

Mr. Malkin is supported by the European Commission CHIL project under contract No. 506909. Mr. Kraft is supported in part by the German Research Foundation (DFG) as part of the SFB 588 on Humanoid Robots.

8. References

[1] B. Clarkson, "Life patterns: structure from wearable sensors," Ph.D. dissertation, MIT, 2002.

[2] D. P. W. Ellis and K. S. Lee, "Minimal-impact audio-based personal archives," in *Proceedings of the First ACM Workshop on Continuous Archiving of Personal Experiences (CAPRE)*, 2004.

[3] "CHIL Project Website," http://chil.server.de.

[4] "SFB Project Website," http://www.sfb588.uni-karlsruhe.de.

[5] D. Gerhard, "Audio signal classification," Simon Fraser University, Tech. Rep., 2000.

[6] N. Oliver, E. Horvitz, and A. Garg, "Layered representations for human activity recognition," in *Proceedings of the International Conference on Multimodal Interfaces*, 2002.

[7] M. J. Reyes-Gomez and D. P. W. Ellis, "Selection, parameter estimation, and discriminative training of hidden markov models for general audio modeling," in *Proceedings of the International Conference on Multimedia and Expo (ICME)*, 2003.

[8] A. Temko and C. Nadeu, "Classification of meeting-room acoustic events with support vector machines and variable-feature-set clustering," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[9] R. Malkin, D. Macho, A. Temko, and C. Nadeu, "First evaluation of acoustic event classification systems in the chil project," Presentation, Joint Workshop on Hands-Free Speech Communication and Microhpone Arrays (HSCMA), 2005.

[10] D. P. W. Ellis and K. S. Lee, "Features for segmenting and classifying long-duration recordings of personal audio," in *Proceedings, Workshop on statistical and perceptual audio processings (SAPA)*, 2004.

[11] A. Kapoor and S. Basu, "The audio epitome: a new representation for modeling and classifying auditory phenomena," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.

[12] H. Yu, "Phase-space representation of speech — revisiting the delta and double-delta deatures," in *Proceedings of the International Conference on Speech and Language Processing (ICSLP)*, 2004.

[13] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. Wiley-Interscience, 2001.

[14] A. J. Bell and T. J. Sejnowski, "Learning the higher-order structure of a natural sound," *Network: Computation in Neural Systems*, no. 7, pp. 261–266, 1996.

[15] M. Lewicki, "Efficient coding of natural sounds," *Nature Neuroscience*, vol. 5, no. 4, pp. 356–363, 2002.

[16] P. Smaragdis, "Redundancy reduction for computational audition, a unifying approach," Ph.D. dissertation, MIT, 2001.

[17] A. J. Bell and T. J. Sejnowksi, "An information-maximisation approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, 1995.

[18] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1483–1492, 1997.

[19] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Prentice Hall, 1993.