ATTENTION BASED TEMPORAL FILTERING OF SENSORY SIGNALS FOR DATA REDUNDANCY REDUCTION

Sofoklis Kakouros[1], *Okko Räsänen*[1], *Unto K. Laine*[1]

[1]Department of Signal Processing and Acoustics, Aalto University, Finland

ABSTRACT

Since modern computational devices are required to store and process increasing amounts of data generated from various sources, efficient algorithms for identification of significant information in the data are becoming essential. Sensory recordings are one example where automatic and continuous storing and processing of large amounts of data is needed. Therefore, algorithms that can alleviate the computational load of the devices and reduce their storage requirements by removing uninformative data are important. In this work we propose a method for data reduction based on theories of human attention. The method detects temporally salient events based on the context in which they occur and retains only those sections of the input signal. The algorithm is tested as a pre-processing stage in a weakly supervised keyword learning experiment where it is shown to significantly improve the quality of the codebooks used in the pattern discovery process.

Index Terms— attention modeling, data redundancy reduction, data compression, machine learning

1. INTRODUCTION

The development of information technology has led to an almost ubiquitous presence of devices which can create, store, and process data. Furthermore, much of this data is not user generated but can be collected automatically by devices with the appropriate sensory equipment. A typical example is a mobile device which is capable of collecting and storing a multitude of sensory data including, for instance, accelerometer, audio, and location data. As the amount of data is becoming increasingly high for a device to process and store, means of reducing and therefore alleviating the computational load for the device, especially given the resource constraints (for instance battery autonomy, processor speed, applications running), need to be considered. The traditional approach to data reduction is by feature extraction and compression where data size is reduced throughout the data and the signal is post-processed in its entirety (common post-processing scenarios for mobile devices include context recognition and activity recognition). Another approach to the data reduction problem is to discard or reduce the resolution of only parts of the signal which do not carry significant information with regard to the system before post processing. In this view, a number of different examples exist (see [1-7]) where the main focus is on utilizing theories from the study of human attention and applying them on computational implementations of data selection models.

The term of attention is frequently used with reference to both input selectivity and capacity constraints. Selectivity, from the perspective of human perception, is reflected in the small fragment of the sensory information which finally reaches our awareness. Capacity constraints, on the other hand, are illuminated by the difficulty of executing multiple tasks simultaneously. These ideas extend to engineering where computational modeling of attention has many areas of application such as machine vision [1,2], audio processing [3,4], data reduction and compression [5,6]. Currently the overall focus of the effort in these areas is on the development of techniques to detect salient features in signals and thus reducing the resolution of signal segments that are deemed non-significant. This typically leads to a spatial, temporal or spatiotemporal reduction in the resolution. For example, in the visual domain, Bruce and Tsotsos [4] proposed a method for saliency computation based on the idea that localized saliency is underpinned by a maximization of the information collected from one's environment. In the auditory domain, Kayser et al. [5] proposed an auditory saliency map in an attempt to describe and understand the process of auditory selectivity. This map can be used in order to extract salient events in natural acoustic scenarios. Finally, regarding data reduction and compression, the main approaches have to do with computation of saliency maps and (i) resolution reduction [5,6], and (ii) signal cropping [7]. The majority of the work, however, is carried out in the domains of image, video, and audio.

This work extends the efforts as presented in [7] for data reduction by utilizing knowledge from human attention theories [8] and memory [9]. Previous work (see e.g., [3-7]) focused primarily on saliency detection based on processed representations of the signal at hand. For example, Kayser et al. [5] used the signal's spectral properties whereas Wrigley and Brown [3] utilized a complex architecture for processing the signal at various different levels (e.g. cochlear filtering, corellogram).

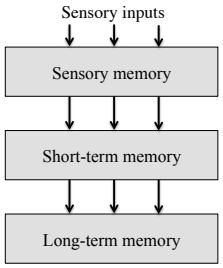

Fig. 1: Multi-store memory model (adapted from Atkinson and Shiffrin [9])

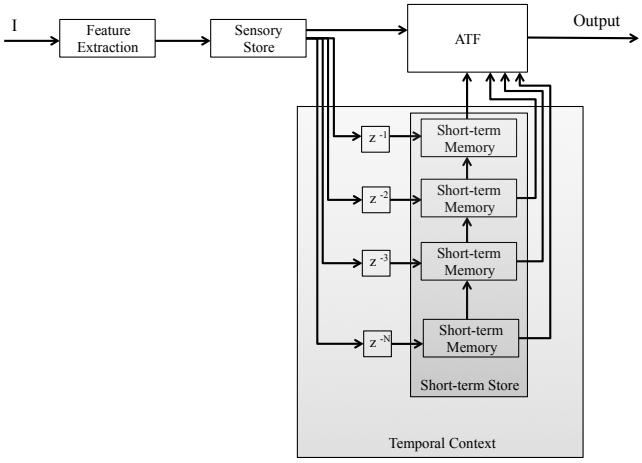

Fig. 2: Overview of ATF structure

In the current work, a novel approach for data reduction by attention based temporal filtering in generic sensory data is presented. The attention temporal filter (ATF) approach differentiates from earlier attempts in that it utilizes the signal properties purely in the time domain. The method and model behind it draws from the early attention theories and combines a human memory model in order to build the method for context-based selection of significant temporal events. The basic idea is that as a signal's properties change over time, the incurred changes carry significance with regard to the temporal context at the time they occurred. Therefore, significant events, and in extent, more temporally salient events, are more likely to carry important information for the system at hand. The proposed method is tested as a pre-processing stage in an unsupervised keyword learning experiment which simulates human infant language acquisition process. The results are compared against a baseline system without any data cleaning.

2. METHODS

The detection of temporally salient events in ATF is performed on the basis of the creation of a *context temporal distance matrix (CTDM)*. The algorithm consists of the following steps: (i) generation and initialization of CTDM, (ii) attenuation of data in CTDM, and (iii) hierarchical clustering and data selection.

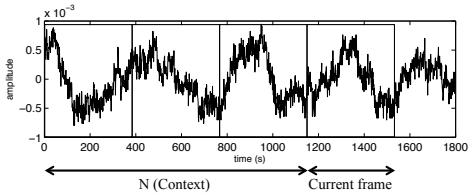

Fig. 3: Structure of context

2.1. Context temporal distance matrix

In ATF, the fundamental element for the processing of the sensory signals is the CTDM. To set the ground for the analysis of CTDM, two basic concepts derived from Atkinson's and Shiffrin's multi-store memory model [9] are essential (see Figure 1), namely: (i) the sensory store (st_{se}) and (ii) the short-term store (st_{st}). The former defines a memory store which is associated with each sensory input and where it holds information for a very short period of time whereas the latter defines a memory store which has also very limited capacity and which holds information for longer periods of time. These components precede the long-term storage of information in humans and therefore data held in them is not necessarily permanently stored. The short-term store can consequently keep information about the context with which a current sensory input can be compared against. Extending this model to the ATF, the sensory store can be defined as the current signal input under examination while the short-term store can be defined as a more extensive collection of previous inputs which stay in the memory buffer in a queue-type manner (first in first out). Therefore, st_{se} holds the current system input while st_{st} holds the system's context (earlier inputs, see Figure 2,3).

Specifically, the signal is divided into frames of duration that depends on the type of the input. For audio input this corresponds to frames of 25 ms without overlap ($t_{st}^{se} = 25ms$). According to [10], st_{st} has the capacity to hold approximately 7±2 individual frames of information. Hence, the context size N can take values between 1 and 9.

$$N = \frac{t_{st}^{st}}{t_{st}^{se}} \Leftrightarrow t_{st}^{st} = N \cdot t_{st}^{se}, N \in \{1,2,\ldots,9\} \quad (1)$$

Therefore, the temporal context for the current input frame will have length N previous frames and duration of t_{st}^{st} ms (see Figure 2). Based on this, CTDM matrix of size NxN is constructed. In this matrix, the first row represents the change between the current frame (f_c) and the previous N frames (f_{c-i}, $1 \le i \le N$). For every new frame that arrives at the system, the context change is recalculated and placed at the first row, therefore pushing the previous, one position lower (e.g. row i to row $i+1$) and pushing row N out of the matrix.

The context change is calculated by taking the Euclidean distances between the current frame and each of the context frames, thereby generating one new first row for CTDM (equations 2, 3). The current input frame in the matrix is

denoted with f_{c^0} while $f_{c^{-i}}$ ($1 \le i \le N-1$) denotes the current inputs of previous context changes. For each frame, one time domain feature is extracted which is used in the calculation of the Euclidean distances. Typical features used are the energy or the average amplitude of the signal.

$$d(f_c, f_{c-1}) = \sqrt{\sum_{i=1}^{n} (f_{c_i} - f_{c-1_i})^2} \quad (2)$$

$$C_{NxN}^{tdm} = \begin{bmatrix} d(f_{c^0}, f_{c^0-1}) & d(f_{c^0}, f_{c^0-2}) & \cdots & d(f_{c^0}, f_{c^0-N}) \\ d(f_{c^{-1}}, f_{c^{-1}-1}) & d(f_{c^{-1}}, f_{c^{-1}-2}) & \cdots & d(f_{c^{-1}}, f_{c^{-1}-N}) \\ \vdots & \vdots & \ddots & \vdots \\ d(f_{c^{-(N-1)}}, f_{c^{-(N-1)}-1}) & d(f_{c^{-(N-1)}}, f_{c^{-(N-1)}-2}) & \cdots & d(f_{c^{-(N-1)}}, f_{c^{-(N-1)}-N}) \end{bmatrix} \quad (3)$$

2.2. Attenuation of data in CTDM

Data stored in CTDM are perceptually weighted in order to account for the attenuation of the short-term memory at different time lags. This is based on the findings of Peterson and Peterson [11] who addressed the problem of information recollection from the short-term store. According to their results, the ability to recall items from short-term memory dropped rapidly over time (approximately as a linear function of time). Therefore, in order to simulate this behavior in CTDM, an attenuation matrix is generated. The elements of the attenuation matrix are calculated based on equation 5 and appear in the form of equation 4. Equation 4 generates weighs for each element in the *N*x*N* matrix and attenuates all other elements except the first (C_{11}). The weighing is performed in a liner manner, where cells in CDTM closer to the first element in the matrix are attenuated less while the ones further away are attenuated the most. The variable *S* in the equation is the attenuation step which is calculated as $S=0.9/N$, where *N* is the size of the context. The mitigated distances in the attenuated CDTM (equation 6) have perceptual significance as frames that are further away from the current may carry less significance in the given context.

$$C_{NxN}^{att} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & \cdots & a_N \\ a_2 & a_2 & \cdots & a_N \\ \vdots & \vdots & \ddots & \vdots \\ a_N & a_N & \cdots & a_N \end{bmatrix} \quad (4)$$

$$a_{ij} = \begin{cases} 1, i = j = 1 \\ 1 - (i-1) \cdot S, j \le i \\ 1 + S \cdot (1-j), j > i \end{cases} \quad (5)$$

$$C^{atdm}{}_{NxN} = C_{NxN}^{att} \circ C_{NxN}^{tdm} \quad (6)$$

2.3. Hierarchical clustering and data selection

Clustering is applied to all individual entries of the attenuated CTDM matrix for each context change (shift of the current frame to the next). The entries are classified into *k* clusters, where *k* is chosen to be close to *N* in order to reflect the potential level of variability in the context. In order to achieve a hierarchical indexing of the elements in CTDM, the *k* clusters are assigned *k* equally spaced centroids between the minimum and maximum values of CDTM. Therefore the produced set of clusters *S* will be an order set with centroids *m* such as:

$$S = \{S_1, S_2, \cdots, S_k\} \quad where \quad m_1 < m_2 < \cdots < m_k$$

This allows the generation of a ranked CTDM with elements that contain indexes of their contextual significance. That is, elements which have low index have low significance in the context while the ones with high index have correspondingly high significance. The selection decision of the current frame in the context is performed on the basis of C_{11}. A typical rule is that if $C_{11} \in S_k$, k=N then the frame is allowed to pass through the filter. Less strict rules can be also applied where the filter will be less sensitive to context changes such as if $C_{11} \in (S_k \cup S_{k-1} \cup S_{k-2})$.

3. EXPERIMENTS

3.1. Experimental setup and evaluation

The performance of the ATF filtering is demonstrated in a weakly supervised word learning experiment from continuous child-directed speech (see [12–14]) using the CM algorithm [12]. In the experiment, the task of the learning algorithm is to discover acoustic patterns (words) in speech that co-occur with contextual labels denoting the keywords present in each utterance. Unlike standard supervised training, the alignment between audio and labels is only available at the utterance level and the relative ordering of keywords is unknown. This type of simulation is often used to study audiovisual associative learning of the human language acquisition process [15].

For data, one speaker (Female-01) from the CAREGIVER Y2 UK corpus is used [16]. The data contains 2397 utterances of continuous English speech with 1-4 keywords occurring in each sentence. In addition, the keywords are surrounded by carrier sentences containing verbs, function words and such. There are a total of 50 unique keywords in the material. In the experiments, the 2000 first utterances are used to train the recognizer for the keywords and testing is performed with the remaining 397 utterances. For each test utterance, the classifier was asked to provide *N* most likely keywords and these hypotheses were then compared against the true *N* words in the annotation. Overall recognition accuracy was measured as the proportion of correct hypotheses, i.e. $N_{\mathrm{CORRECT}}/N_{\mathrm{TOT}}$ over all 397 sentences in the test set.

The speech data was pre-processed by first applying the ATF to the original audio waveforms (f_s = 16000 Hz) and extracting 39 dimensional MFCC features (13 static, 13 ΔMFCC, and 13 ΔΔMFCC coefficients) with 25–ms window length using 10 ms shifts. Then the MFCC frames filtered away by the ATF were discarded but the timing information of the remaining frames was maintained. Finally, the MFCCs were vector quantized (VQ) by creating a codebook of size N_A using k-means clustering on 10000

randomly chosen features from the training data and then quantizing all frames using the codebook. Four different codebook sizes (N_A = 32, 64, 128 and 256) were studied in the experiments.

As the classifier, we used the concept matrix (CM) algorithm [12] designed for the current type of weakly supervised learning from discrete sequential data. CM finds the correspondence between acoustic patterns (VQ sequences of words) and the contextual labels by modeling the VQ data as a mixture of bi-grams measured from different temporal distances and finding the maximum likelihood solution that a set of lagged bi-grams occurs during a specific keyword.

During recognition, given input sequence X, the total activation of a keyword c is measured as

$$A(c \mid X) = \frac{1}{LK} \sum_{t=K+1}^{L} \sum_{k=1}^{|\mathbf{k}|} P(c \mid X(t), X(t-k), k) \qquad (7)$$

where L is the length of the signal, k is the lag at which bi-gram is computed and $|\mathbf{k}|$ is the total number of lags ($\mathbf{k}$ = {1, 2, …, 20} in the current work). The CM-model P($c|X$) on the right-hand side of the equation was computed according to [12]. The N most likely keywords were chosen as the hypotheses for the input X (see above).

In order to compare the effect of ATF to the baseline performance, three experimental conditions were studied: baseline keyword learning result without ATF ("baseline"), filtering all training and testing data with the ATF ("F-ATF"), and filtering the data for VQ codebook generation but training and testing the classifier itself with full-length speech signals ("VQ-ATF"). The experiment was repeated eight times for each condition in order to measure the average performance across the trials.

3.2. Results

Figure 4 shows an example of applying ATF filtering to a speech signal. As can be observed, it efficiently filters out the silent portions of the signal preceding and following the utterance, but also some stationary parts of the signal, effectively acting as a combined onset- and voice activity detector. Figure 5 shows the keyword recognition performance as a function of codebook size for the three experimental conditions. The vertical bars denote the standard deviation across trials.

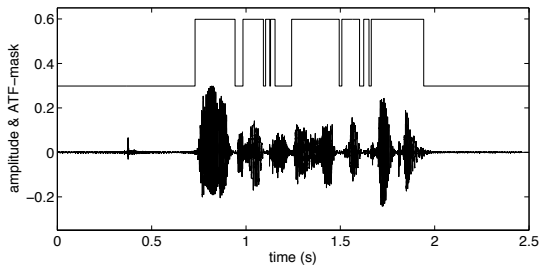

Fig. 4: An example of ATF filtering output for utterance "*Mommy takes the happy cookie*". The binary mask shows passed samples (high value) and filtered samples (low value).

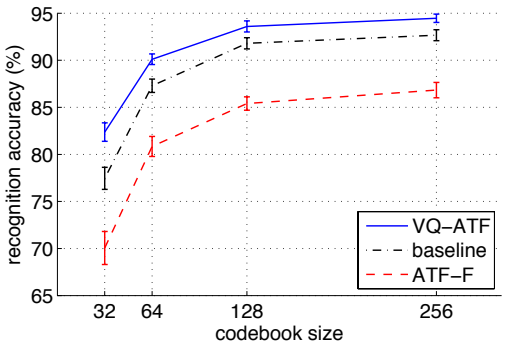

Fig. 5: Keyword recognition accuracies as a function of codebook sizes for four different codebook sizes.

As can be observed, the VQ-ATF performs the best across all variants. On average, VQ-ATF leads to 29% reduction in word error rate in comparison to the baseline system ($p << 0.01$, paired t-test) across all codebook sizes. Also, the standard deviation of performance across different iterations is smaller for VQ-ATF than the baseline system (without any filtering). In contrast, the F-ATF system that uses only a subset of the frames of each signal data performs at the worst level. This is not surprising, however, since the data reduction is notable (approx. 57.7%) and mainly transitions and onsets of the stationary periods are maintained in the signal.

In general, it seems that the application of ATF to the data used in codebook generation enhances the quality of the codebooks by allowing more principled selection of feature frames to the clustering process. This suggests that it may also be suitable for other systems susceptible to proper selection of representative feature frames such as the k-Nearest Neighbors classifier or even Gaussian Mixture Models. However, according to the findings, it may not be suitable as a front-end for sequential classifiers such as CMs or Hidden-Markov Models during the recognition stage due to the fact that typically redundant or even low-quality features are more beneficial in the recognition process than no features at all.

4. CONCLUSIONS

In this work, a novel approach for data redundancy reduction based on the detection of temporally salient events was presented. The proposed method is based on theories of attention and memory and performs a filtering operation on sensory signals in order to select temporally important inputs in the defined context. Based on the experiments, the ATF approach significantly reduced the error rate as it improved the feature selection for the clustering of codebooks for pattern recognition. In future work, the ATF will be tested using different sensory signals and on other recognition tasks (such as activity detection).

5. ACKNOWLEDGEMENTS

This research was funded by Tekes program From Data to Intelligence (D2I).

6. REFERENCES

[1] Itti, L. and Koch, C., "Computational Modeling of Visual Attention," *Nature Reviews Neuroscience*, Vol. 2, pp. 194–203, 2001.

[2] Bruce, N. D. and Tsotsos, J. K., "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, Vol. 9(3), pp. 1–24, 2009.

[3] Wrigley, S. N. & Brown G. J., "A Computational Model of Auditory Selective Attention," *IEEE Transactions on Neural Networks*, Vol. 15(5), pp. 1151–1163, 2004.

[4] Kayser, C., Petkov, C., Lippert, M. and Logothetis, N. K., "Mechanisms for allocating auditory attention: An auditory saliency map," *Current Biology*, Vol. 15(21), pp. 1943–1947, 2005.

[5] Li, Z., Qin, S. and Itti, L., "Visual attention guided bit allocation in video compression," *Image and Vision Computing*, 29(1), pp. 1–14, 2011.

[6] Lee, J., De Simone, F. and Ebrahimi, T., "Efficient video coding based on audio-visual focus of attention," *Journal of Visual Communication and Image Representation*, Vol. 22(8), pp. 704–711, 2010.

[7] Mancas M., Beul, D.D., Riche, N., and Siebert, X., "Human Attention Modelization and Data Reduction," In: Intech, ed. Intech., *Video Compression*, 2012.

[8] Broadbent, D. E., "Perception and Communication," New York: Pergamon, 1958.

[9] Atkinson, R.C., and Shiffrin, R.M., "Chapter: Human memory: A proposed system and its control processes," In Spence, K.W., and Spence, J.T., The psychology of learning and motivation (Vol. 2), pp. 89–195, New York: Academic Press, 1968.

[10] Miller, G.A., "The magic number seven, plus or minus two: some limits on our capacity for processing of information," *Psychological Review*, Vol. 63, pp. 81–93, 1956.

[11] Peterson, L.R. and Peterson, M.J., "Short-term retention of individual items," *Journal of Experimental Psychology*, Vol. 58(3), pp. 193–198, 1959.

[12] Räsänen O., and Laine U., "A method for noise-robust context-aware pattern discovery and recognition from categorical sequences," *Pattern Recognition*, Vol. 45, pp. 606–616, 2012.

[13] ten Bosch, L., Van hamme, H., Boves, L., Moore, R.K., "A computational model of language acquisition: the emergence of words," *Fundamenta Informaticae*, Vol. 90, pp. 229–249, 2009.

[14] Driesen, J. and Van hamme, H., "Supervised input space scaling for non-negative matrix factorization," *Signal Processing*, Vol. 92, pp. 1864–1874, 2012.

[15] Räsänen O.: "Computational modeling of phonetic and lexical learning in early language acquisition: existing models and future directions," *Speech Communication*, Vol. 54, pp. 975–997, 2012.

[16] Altosaar, T., ten Bosch, L., Aimetti, G., Koniaris, C., Demuynck, K., & van den Heuvel, H., "A Speech Corpus for Modeling Language Acquisition: CAREGIVER," *Proceedings of the International Conference on Language Resources and Evaluation* (LREC), Malta, pp. 1062–1068, 2010.