MAX-PRODUCT DYNAMICAL SYSTEMS AND APPLICATIONS TO AUDIO-VISUAL SALIENT EVENT DETECTION IN VIDEOS

Petros Maragos and Petros Koutras

School of ECE, National Technical University of Athens, 15773 Athens, Greece
maragos@cs.ntua.gr, pkoutras@cs.ntua.gr

ABSTRACT

This paper introduces a theory for max-product systems by analyzing them as discrete-time nonlinear dynamical systems that obey a superposition of a weighted maximum type and evolve on nonlinear spaces which we call complete weighted lattices. Special cases of such systems have found applications in speech recognition as weighted finite-state transducers and in belief propagation on graphical models. Our theoretical approach establishes their representation in state and input-output spaces using monotone lattice operators, finds analytically their state and output responses using nonlinear convolutions, studies their stability, and provides optimal solutions to solving max-product matrix equations. Further, we apply these systems to extend the Viterbi algorithm in HMMs by adding control inputs and model cognitive processes such as detecting audio and visual salient events in multimodal video streams, which shows good performance as compared to human attention.

Index Terms— nonlinear systems, multimedia signal processing, lattices, minimax algebra, event detection, cognitive modeling.

1. INTRODUCTION AND SUMMARY

Several successful algorithms in pattern recognition and machine learning are based on a max-product arithmetic. Examples include speech recognition using weighted finite-state transducers (WFSTs) [32, 20], belief propagation in probabilistic graphical models [3, 40], and the maximum approximation used by the Viterbi decoding algorithm for likelihood scores during state estimation [33]. Further in signal processing and control there are several established areas using max/min superpositions and related operations of signals or vectors; examples include (i) the max-plus convolution (a.k.a. dilation) in morphological signal/image processing [18, 28, 36, 38] convex analysis [26, 35] and optimization [1], (ii) the minimax algebra used in scheduling [12], and (iii) the max-plus control in discrete-event dynamical systems [11, 23, 9]. Further, in multimodal signal processing for cognition modeling, which has been a main motivation for this work, several psychophysical and computational experiments indicate that the superposition of sensory signals or cognitive states seems to be better modeled using max or min rules, possibly weighted. Such an example is the recent work [15] on attention-based multimodal video summarization where a (possibly weighted) min/max fusion of features from the audio and visual signal channels and of salient events from various modalities seems to outperform linear fusion schemes. Finally, the sensory-semantic integration problem in multimedia signal processing requires fusion of two different continuous modalities (audio and vision) with discrete language symbols and semantics extracted from text. Similarly, in control and robotics there are efforts to develop hybrid systems that can model interactions between heterogeneous information streams like continuous inputs and symbolic strings [5]. In both of these applications we need models where the computations among modalities/states can handle both real numbers and Boolean variables; this is possible using max/min rules.

Motivated by the above multimodal signal processing problems, in this paper we develop some theoretical tools for the representation and analysis of nonlinear systems whose dynamics evolve based on the following **state-space max-product model**:

$$\begin{array}{rcl} \boldsymbol{x}(t) & = & \boldsymbol{A}(t) \boxtimes \boldsymbol{x}(t-1) \ \vee \ \boldsymbol{B}(t) \boxtimes \boldsymbol{u}(t) \\ \boldsymbol{y}(t) & = & \boldsymbol{C}(t) \boxtimes \boldsymbol{x}(t) \ \vee \ \boldsymbol{D}(t) \boxtimes \boldsymbol{u}(t) \end{array} \tag{1}$$

where t denotes a discrete time index, $\vee$ denotes maximum, $\boldsymbol{x}(t)$ is an evolving state vector, $\boldsymbol{u}(t)$ is the input signal (scalar or vector), $\boldsymbol{y}(t)$ is an output signal (scalar or vector), and $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ are appropriately sized matrices. $\boxtimes$ denotes the following nonlinear matrix product with max-product operations:

$$\boldsymbol{P} = \boldsymbol{Q} \boxtimes \boldsymbol{R}, \quad p_{ij} = \bigvee_k q_{ik} \times r_{kj} \tag{2}$$

The state equations (1) are written for the case of time-varying coefficients. If the matrices are constant and under zero-initial conditions, the input-output relationship of (1) can be described by a **max-product convolution**:

$$y(t) = (h \otimes u)(t) = \bigvee_k u(k)h(t-k) \tag{3}$$

where h is the system's impulse response. By replacing maximum ($\vee$) with minimum ($\wedge$) in (1) and (3) we can also obtain a *dual* model that describes the dynamics of min-product systems.

Compare the above with linear systems [4, 6, 22, 17], which deal with linear maps: $\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{x}(t-1) + \boldsymbol{B}\boldsymbol{u}(t)$ and $\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}\boldsymbol{u}(t)$. There, all the matrix-vector products and signal convolutions are linear, based on a sum-of-products arithmetic.

A max-product system is a special case of more general systems, studied in detail in [30], whose algebra is based on maximum of $\star$ operations. Examples of 'multiplication' $\star$ include the sum and the product, but $\star$ may be only a semigroup operation. The resulting algebras include the *max-plus algebra* $(\mathbb{R} \cup \{-\infty\}, \max, +)$ used in scheduling and operations research [12], discrete-event dynamical systems [10, 11, 8, 9], automated manufacturing [23, 24, 13] and max-plus control [10, 16, 7]; the min-plus algebra or else known as *tropical semiring* $(\mathbb{R} \cup \{+\infty\}, \min, +)$ used in shortest paths on networks [12] and in natural language processing [32, 20]; the *fuzzy*

This research work was supported by the project COGNIMUSE which is implemented under the ARISTEIA Action of the Operational Program Education and Lifelong Learning and is co-funded by the European Social Fund and Greek National Resources.

978-1-4673-6997-8/15/$31.00 ©2015 IEEE

logic semiring $([0,1],\vee,T)$ with statistical T-norms playing the role of fuzzy intersection used in fuzzy automata and neural nets [25, 21], and fuzzy dynamical systems [31].

Our Contributions. (1) Developed a theory for max-product systems analyzing both their dynamics in state-space and their input-output convolutional representation by using a new and powerful class of underlying spaces, the *complete weighted lattices (CWLs)*. The detailed theory of CWLs is developed in [29, 30] to which we refer the reader for all proofs. (2) Derived analytic formulae for computing the state and output responses of max-product systems as well as for finding their input-output max-product convolutions, represented in both cases via lattice monotone operators in adjunction pairs. Further, use the latter to generate optimal solutions for solving max-product equations $\boldsymbol{A} \boxtimes \boldsymbol{x} = \boldsymbol{b}$. (3) Studied various control-theoretic issues of max-product systems. (4) Developed applications of max-product systems that extend the Viterbi algorithm of hidden Markov models (HMMs) to cases with control inputs and can estimate the saliencies of audio-visual events in multimodal videos with good performance as compared to human attention.

2. BACKGROUND ON LATTICES AND OPERATORS

The background material in this section follows [2], [37], [19], [18] and [29]. A partially-ordered set, briefly **poset** $(\mathcal{P},\leq)$, is a set $\mathcal{P}$ in which a *partial ordering* $\leq$ is defined. If the ordering $\leq$ is total, then we have a *chain*. A **lattice** is a poset $(\mathcal{L},\leq)$ any two of whose elements have a *supremum* (a.k.a. least upper bound), denoted by $X \vee Y$, and an *infimum* (a.k.a. greatest lower bound), denoted by $X \wedge Y$. We often denote the lattice structure by $(\mathcal{L},\vee,\wedge)$. A lattice $\mathcal{L}$ is *complete* if each of its (finite or infinite) subsets has a supremum and an infimum in $\mathcal{L}$.

Duality: In any lattice $\mathcal{L}$, by replacing the partial ordering $\leq$ with its dual $\leq'$ and by interchanging the roles of the supremum and infimum, we can form a new lattice called the *dual lattice* and often denoted by $\mathcal{L}'$. To every definition, property and statement that applies to $\mathcal{L}$ there also corresponds a dual one that applies to $\mathcal{L}'$.

Examples of Complete Lattices: (a) The chain of extended real numbers $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty,+\infty\}$ equipped with the natural order $\leq$. (b) The *power set* $\mathcal{P}(E) = \{X : X \subseteq E\}$ of an arbitrary set E equipped with the partial order of set inclusion where the supremum and infimum are the set union and intersection. (c) *Function Lattices*: The set of discrete-time signals $f : \mathbb{Z} \to \overline{\mathbb{R}}$ equipped with the *pointwise* ordering $\leq$, supremum and infimum of $\overline{\mathbb{R}}$.

Increasing Operators: Given two operators ψ and ϕ on a complete lattice $\mathcal{L}$ we can define *pointwise* a partial ordering $\leq$ between them, their supremum $(\psi \vee \phi)$ and infimum $(\psi \wedge \phi)$. Further, we define the composition of two operators as an operator product: $\psi\phi(X) \triangleq \psi(\phi(X))$; special cases are the operator powers $\psi^n = \psi\psi^{n-1}$. Some useful types and properties of lattice operators ψ include the following: (i) identity: $\mathbf{id}(X) = X \quad \forall X \in \mathcal{L}$. (ii) extensive: $\psi \geq \mathbf{id}$. (iii) anti-extensive: $\psi \leq \mathbf{id}$. (iv) idempotent: $\psi^2 = \psi$.

A lattice operator ψ is called *increasing* if it is order-preserving, i.e. $X \leq Y \Longrightarrow \psi(X) \leq \psi(Y)$. Four important types of increasing operators are the following:

δ is *dilation* iff $\delta(\bigvee_i X_i) = \bigvee_i \delta(X_i)$
ε is *erosion* iff $\varepsilon(\bigwedge_i X_i) = \bigwedge_i \varepsilon(X_i)$
α is *opening* iff α is increasing, idempotent & anti-extensive
β is *closing* iff β is increasing, idempotent & extensive

The four above types of lattice operators were originally defined in [37, 18] as generalizations of the corresponding standard morphological image operators.

Dilations and erosions come in pairs as the following concept reveals. The pair (ε,δ) of two operators δ and ε on a complete lattice $\mathcal{L}$ is called an **adjunction** on $\mathcal{L}$ if

$$\delta(X) \leq Y \Longleftrightarrow X \leq \varepsilon(Y) \quad \forall X,Y \in \mathcal{L} \tag{4}$$

In any adjunction (ε,δ), ε is called the *adjoint erosion* of δ, whereas δ is the *adjoint dilation* of ε. There is a one-to-one correspondence between the two operators of an adjunction, since, given a dilation δ, there is a unique erosion

$$\varepsilon(Y) = \bigvee\{X \in \mathcal{L} : \delta(X) \leq Y\} \tag{5}$$

such that (ε,δ) is adjunction, and vice-versa.

From the composition of the erosion and dilation of any adjunction (ε,δ) we can generate an opening $\alpha = \delta\varepsilon$; since α is an opening, we have $\alpha(f) \leq f$ and $\alpha^2 = \alpha$. Dually, any adjunction can also generate a closing $\beta = \varepsilon\delta$. Both of these are special cases of morphological filters in [37, 18], a.k.a. *lattice projections* [29], since they are increasing and idempotent.

3. THEORY OF MAX-PRODUCT SYSTEMS

3.1. Weighted Lattices of Vectors and Signals

All elements of the vectors, matrices, or signals involved in the description of max-product systems take their values from the set $\mathcal{K} = [0,\infty]$ of nonnegative extended reals. We equip $\mathcal{K}$ with the following scalar operations: (A) the standard maximum or supremum $\vee$ on $\overline{\mathbb{R}}$, which plays the role of a generalized *'addition'*. (A$'$) the standard minimum or infimum $\wedge$ on $\overline{\mathbb{R}}$. It plays the role of a generalized *'dual addition'* . (M) the *multiplication* $\times$ extended over $[0,\infty]$ which has 1 as its identity and 0 as its null element, and distributes over any supremum. (M$'$) a *'dual multiplication'* $\times'$ which has ∞ as null element, distributes over any infimum and coincides with $\times$ on $(0,\infty)$. The four above operations make $\mathcal{K}$ an algebraic structure called *clodum* (complete lattice-ordered double monoid) [27, 29]. We can also define a *conjugation* operation mapping bijectively each element a to its *conjugate* element $\overline{a} = 1/a = a^{-1}$. This interchanges suprema with infima; further $\overline{a \times b} = a^{-1} \times' b^{-1}$. In $[0,\infty]$ the $\times$ and $\times'$ operations coincide in all cases with only one exception, the multiplication of 0 with ∞. Thus, henceforth we shall use only one multiplication $(\times)$ and remember that the case $0 \times \infty$ will have value 0 (resp. ∞) if it is combined with other terms via a supremum (resp. infimum).

Consider the set $\mathcal{W}$ consisting of all nonnegative functions $F : E \to \mathcal{K}$ defined on an arbitrary nonempty set E and taking values in the clodum $\mathcal{K} = [0,\infty]$. If we extend *pointwise* the supremum $(F \vee G)$, infimum $(F \wedge G)$ and scalar multiplication $(a \times F)$ for functions $F, G \in \mathcal{W}$ and scalars $a \in \mathcal{K}$, the set $\mathcal{W}$ becomes a *complete weighted lattice (CWL)* over $\mathcal{K}$. We can also have conjugation of functions by defining $\overline{F}(t) = 1/F(t)$. The axioms of CWLs bear a remarkable conceptual similarity with those of linear spaces as analyzed in in our recent work [29, 30]. We focus on two special cases: (i) If $E = \{1,2,...,n\}$, then $\mathcal{W}$ becomes the set of all n-dimensional vectors with elements from $\mathcal{K}$. (ii) If $E = \mathbb{Z}$, then $\mathcal{W}$ becomes the set of all discrete-time signals with values from $\mathcal{K}$.

On linear spaces, a linear system Γ obeys *linear* superposition:

$$\Gamma(\sum_i a_i F_i) = \sum_i a_i \Gamma(F_i) \tag{6}$$

On a CWL the conceptually analogous superposition would be to have systems δ that obey a max-product superposition:

$$\delta(\bigvee_i c_i F_i) = \bigvee_i c_i \delta(F_i), \tag{7}$$

This means that δ is both a dilation and invariant to vertical scalings (in short V-scalings) of signals $F(t) \mapsto aF(t)$. We call δ a *dilation V-scaling invariant (DVI)* system.

CWL of vectors: Consider now the CWL vector space $\mathcal{W} = \mathcal{K}^n$, equipped with the pointwise partial ordering $\boldsymbol{x} \leq \boldsymbol{y}$, supremum $\boldsymbol{x} \vee \boldsymbol{y} = [x_i \vee y_i]$, infimum $\boldsymbol{x} \wedge \boldsymbol{y} = [x_i \wedge y_i]$, and scalar multiplications of vectors . On finite-dimensional linear vector spaces a vector map is linear iff it can be represented as a linear product between the system's matrix and the input vector. Similarly, we have shown that on the CWL $\mathcal{W}$ a map is DVI iff it can be represented as the max-product between the input vector $\boldsymbol{x}$ and the matrix $\boldsymbol{M} = [m_{ij}]$ with $m_{ij} = \{\delta(\boldsymbol{v}_j)\}_i$, where $\boldsymbol{v}_j$ are basis vectors. This map is a vector dilation $\delta_{\boldsymbol{M}}(\boldsymbol{x}) = \boldsymbol{M} \boxtimes \boldsymbol{x}$. Its adjoint vector erosion, so that (ε, δ) is an adjunction, can be shown to equal [30]

$$\varepsilon(\boldsymbol{y}) = \boldsymbol{M}^* \boxtimes' \boldsymbol{y}, \quad \boldsymbol{M}^* \triangleq \overline{\boldsymbol{M}}^T, \tag{8}$$

where $\boldsymbol{M}^* \triangleq [m_{ji}^{-1}]$ is the *adjoint matrix* of $\boldsymbol{M} = [m_{ij}]$, and $\boxtimes'$ denotes the *matrix min-product*; namely, $\boldsymbol{P} = \boldsymbol{Q} \boxtimes' \boldsymbol{R}$ with $p_{ij} = \bigwedge_k q_{ik} r_{kj}$. This adjunction helps us solve **max-product equations**:

$$\boldsymbol{A} \boxtimes \boldsymbol{x} = \boldsymbol{b} \tag{9}$$

Often (9) does not have an exact solution, in which case we can find an optimum approximate solution by solving the following constrained minimization problem:

$$\begin{array}{c} \text{Minimize } \ ||\boldsymbol{A} \boxtimes \boldsymbol{x} - \boldsymbol{b}|| \\ \text{subject to } \ \boldsymbol{A} \boxtimes \boldsymbol{x} \leq \boldsymbol{b} \end{array} \tag{10}$$

where $||\cdot||$ is either the ℓ_∞ or the ℓ_1 norm.

Theorem 1 *(a) The vector $\hat{\boldsymbol{x}} = \boldsymbol{A}^* \boxtimes' \boldsymbol{b}$ is a solution to (10).*
(b) If Eq. (9) has a solution, then $\hat{\boldsymbol{x}}$ is its greatest solution.

Our method for solving (10) is to consider vectors $\boldsymbol{x}$ that are *sub-solutions* in the sense that $A \boxtimes \boldsymbol{x} \leq \boldsymbol{b}$ and find the greatest such sub-solution using adjunctions. The set of sub-solutions forms a semigroup under vector $\vee$ whose supremum equals $\hat{\boldsymbol{x}}$, which yields either the greatest exact solution of (9) or an optimum approximate solution in the sense of (10). This adjunction-based solution creates a lattice projection via the opening $\delta(\varepsilon(\boldsymbol{b})) \leq \boldsymbol{b}$ that best approximates $\boldsymbol{b}$ from below.

CWL of signals: Consider the set $\mathcal{W}$ of all discrete-time signals $f : \mathbb{Z} \to \mathcal{K}$ with values from $\mathcal{K} = [0, \infty]$. Equipped with pointwise supremum $\vee$ and infimum $\wedge$, and pointwise scalar multiplications, this becomes a complete weighted lattice. The signal translations are the operators $\mathcal{T}_{k,v}(f)(t) = vf(t-k)$. A signal operator on $\mathcal{W}$ is called *translation invariant* iff it commutes with any such translation. This translation-invariance contains both a vertical translation and a horizontal translation which is the well-known *time-invariance*. Now, if $q(t)$ is the *impulse*, equal to 1 at $t = 0$ and 0 elsewhere, every signal f can be represented as a supremum of translated impulses

$$f(t) = \bigvee_k f(k) q(t-k) \tag{11}$$

Consider now operators Δ on $\mathcal{W}$ that are dilations and translation-invariant in the above sense. Then, Δ is both DVI in the sense of (7) and time-invariant. We call such operators **dilation translation-invariant (DTI)** systems. Applying Δ to an input signal f decomposed as in (11) yields the output as the max-product convolution $\otimes$ of the input with the system's impulse response $h = \Delta(q)$:

$$\Delta(f)(t) = (f \otimes h)(t) = \bigvee_k f(k) h(t-k) \tag{12}$$

Theorem 2 *A signal operator Δ is a DTI system iff it can be represented as the max-product convolution of the input signal with the system's impulse response $h = \Delta(q)$.*

3.2. State and Output Responses

Based on the state-space model of a max-product dynamical system (1), we can compactly express its state response and output response if we know its *transition matrix*:

$$\boldsymbol{\Phi}(t_2, t_1) \triangleq \begin{cases} \boldsymbol{A}(t_2) \boxtimes \cdots \boxtimes \boldsymbol{A}(t_1 + 1) & \text{if} \quad t_2 > t_1 \\ \boldsymbol{I}_n & \text{if} \quad t_2 = t_1 \end{cases} \tag{13}$$

for $t_2 \geq t_1$, where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. By using induction on (1), the state and output responses of the time-varying nonhomogeneous system can be found, for $t \geq 0$,

$$\boldsymbol{x}(t) \;=\; \boldsymbol{\Phi}(t,0) \boxtimes \boldsymbol{x}(0) \vee \left(\bigvee_{i=1}^{t} \boldsymbol{\Phi}(t,i) \boxtimes \boldsymbol{B}(i) \boxtimes \boldsymbol{u}(i) \right) \tag{14}$$

$$\begin{aligned} \boldsymbol{y}(t) \;=\; & \boldsymbol{C}(t) \boxtimes \Phi(t,0) \boxtimes \boldsymbol{x}(0) \ \vee \ \boldsymbol{D}(t) \boxtimes \boldsymbol{u}(t) \\ & \vee \left(\bigvee_{i=1}^{t} \boldsymbol{C}(t) \boxtimes \boldsymbol{\Phi}(t,i) \boxtimes \boldsymbol{B}(i) \boxtimes \boldsymbol{u}(i) \right) \end{aligned} \tag{15}$$

The zero-state part of $\boldsymbol{y}$ is a *time-varying max-product convolution*. If matrices $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ are constant, the state equations become:

$$\begin{array}{rcl} \boldsymbol{x}(t) & = & \boldsymbol{A} \boxtimes \boldsymbol{x}(t-1) \ \vee \ \boldsymbol{B} \boxtimes \boldsymbol{u}(t) \\ \boldsymbol{y}(t) & = & \boldsymbol{C} \boxtimes \boldsymbol{x}(t) \ \vee \ \boldsymbol{D} \boxtimes \boldsymbol{u}(t) \end{array} \tag{16}$$

and $\Phi(t_2, t_1) = \boldsymbol{A}^{(t_2 - t_1)}$, where $A^{(t)}$ denotes the t-fold max-product of $\boldsymbol{A}$ with itself. By representing the matrix-vector max-product as a dilation operator $\boldsymbol{x} \mapsto \delta_{\boldsymbol{A}}(\boldsymbol{x}) = \boldsymbol{A} \boxtimes \boldsymbol{x}$, the solutions of the *constant-matrix state equations* become

$$\boldsymbol{x}(t) \;=\; \delta_{\boldsymbol{A}}^t[\boldsymbol{x}(0)] \ \vee \ \left(\bigvee\nolimits_{i=1}^{t} \delta_{\boldsymbol{A}}^{t-i} \delta_{\boldsymbol{B}}[\boldsymbol{u}(i)] \right) \tag{17}$$

$$\boldsymbol{y}(t) \;=\; \underbrace{\delta_{\boldsymbol{C}} \delta_{\boldsymbol{A}}^t[\boldsymbol{x}(0)]}_{\text{zero-input resp.}} \ \vee \ \underbrace{\left(\bigvee\nolimits_{i=1}^{t} \delta_{\boldsymbol{C}} \delta_{\boldsymbol{A}}^{t-i} \delta_{\boldsymbol{B}}[\boldsymbol{u}(i)] \right) \vee \delta_{\boldsymbol{D}}[\boldsymbol{u}(t)]}_{\boldsymbol{y}_{zs}(t) \triangleq \text{zero-state resp.}}$$

Thus, the output response is found to consist of two parts: (i) the zero-input response which is due only to the initial conditions $\boldsymbol{x}(0)$ and assumes a zero input, and (ii) the zero-state response which is due only to the input $\boldsymbol{u}(t)$ and assumes zero initial conditions $\boldsymbol{x}(0)$.

For *single-input single-output* systems the mapping $u(t) \mapsto y_{zs}(t)$ can be viewed as a translation invariant dilation system Δ. Hence, the zero-state response can be found as the max-product convolution of the input with the system's impulse response $h = \Delta(q)$. The latter can be found from the general output by setting initial conditions $\boldsymbol{x}(0) = \boldsymbol{0}$ and the input $u(t) = q(t)$:

$$h(t) = \begin{cases} D, & t = 0 \\ \boldsymbol{C} \boxtimes \boldsymbol{A}^{(t)} \boxtimes \boldsymbol{B}, & t \geq 1 \end{cases} \tag{18}$$

The previous results allowed us to address and solve in [30] various important control-theoretic problems for max-product systems, such as their stability, controllability and observability. We outline next the stability result. A useful bound for signals $f(t)$ processed by such systems is their supremal value $\bigvee_t f(t)$. We call max-product systems bounded-input bounded-output (BIBO) *stable* iff an upper bounded input yields an upper bounded output, i.e. if

$$\bigvee_t u(t) < \infty \Longrightarrow \bigvee_t y(t) < \infty \tag{19}$$

Since all signals involved are nonnegative, the above definition of sup-stability coincides with their absolute stability.

Theorem 3 *Consider a DTI system Δ and let $h = \Delta(q)$ be its impulse response. Then: (a) The system is causal iff $h(t) = 0$ for all $t < 0$. (b) The system is BIBO stable iff $\bigvee_t h(t) < \infty$.*

4. HMMS EXTENSIONS AND APPLICATIONS TO DETECTING MULTIMODAL SALIENCIES

Assume a video sequence of audio-visual events each to be scored with some degree of saliency in $[0, 1]$ where 'saliency' is some bottom-up low-level sensory form of attention by a human watching this video. The states x_1, x_2, x_3, x_4 represent time-evolving mono- or multi-modal saliencies, where 1=audio, 2=visual, 3=audiovisual, and 4=non-salient. Peaks in these saliency trajectories signify important events, which can be automatically detected and produce video summaries that agree well with human attention [15]. The following state equations are a possible time-varying max-product dynamical model we propose for the evolution of these saliencies:

$$x_i(t) = \left(\bigvee_{j=1}^{4} a_{ij} x_j(t-1) \right) \star p_i(t) \vee \left(\bigvee_{j=1}^{4} b_{ij} u_j(t) \right) \qquad (20)$$

for state $i = 1, 2, 3, 4$. The constants a_{ij} represent state transitions probabilities and $p_i(t)$ denotes the probability of state $x_i(t)$ being salient based on observed measurable low-level feature vectors $\boldsymbol{o}_t$. We assume that the parameters a_{ij} and $p_i(t)$ are given. The operation $\star$ must distribute over $\vee$ and can be a product, min or max.

Assume first that $\star$ is the product. Given a time sequence of observations $(\boldsymbol{o}_0, \boldsymbol{o}_1, ..., \boldsymbol{o}_t)$ one can fit HMMs to these data using maximum likelihood [34]. Then, the first term in the RHS of (20) models the evolution of the Viterbi dynamic programming (DP) algorithm used in automatic speech recognition with HMMs for optimal state estimation, if we initialize at $t = 0$ the four states by setting $x_i(0) = \pi_i p_i(0)$ where π_i denotes the probability of the system being at the ith state at $t = 0$. For example, if the inputs $u_i(t)$ are all zero, then the single output $y(t) = \bigvee_i x_i(t)$ computes the Viterbi score, which is the probability for having observed the data $(\boldsymbol{o}_0, ..., \boldsymbol{o}_t)$ and the HMM having passed through the optimum state sequence (that maximizes this probability). Our system (20) is more general than the Viterbi algorithm from which it differs in the following aspects: 1) we have the probability-like signals $u_i(t)$ which can act as *control inputs* coming possibly from higher-level events (e.g. detected human faces, presence of speech in the audio, or other semantics). 2) the outputs of the dynamical system can be various min-max combinations of the saliency states of various modalities. 3) the operation $\star$ may be different than the product (which makes the system an HMM if the inputs are zero). For example, it can be a minimum or a maximum.

In our experiments, for estimating the observation data probabilities $p_i(t)$ we have followed two different approaches. In the first, we fitted Gaussian mixture models (GMMs) to audio and visual feature vectors extracted from the video data at each frame t. In the other, we used bottom-up likelihoods by fusing saliencies of the audio and visual streams measured from monomodal cues as in [15]. We have also used high-level control inputs, i.e. automatic face detection [39] and speech activity detection (VAD) [14]. In the case of GMMs we estimated the state transition probabilities a_{ij} using the EM algorithm on some training data from movie videos. In the case of bottom-up likelihoods, the probabilities a_{ij} were set equal to 1/4 plus a penalty at the diagonal elements a_{ii}. For the salient event detection we keep the best state path (the state sequence that has the highest probability) and compare it with human annotations from the movie video. If a frame is annotated with N-labels (e.g. "Audio" and "Audio-Visual"), we search in the N-best state paths. In Table 1 we present our evaluation results on a movie video ('Gladiator') from the MovSum database [15]. We also see the average performance over six movies from various film genres. Our results using the max-product dynamical system are encouraging as they can estimate monomodal or multimodal audio-visual salient events more accurately than GMMs or the bottom-up feature-based likelihoods and can improve with higher-level control inputs. They also outperform HMMs. In Fig. 1 we see an example of our system evolution. Note that in most cases the human-annotated salient events are included in the best state paths found by our system.

	GMM Likelihoods							Bottom-Up (BU) Likelihoods						
	GMM	HMM Variant			MPDS			BU	HMM Variant			MPDS		
State		Prod.	Min	Max	Prod.	Min	Max		Prod.	Min	Max	Prod.	Min	Max
A	65	68	68	64	76	69	67	24	24	34	26	63	71	74
V	50	52	52	45	57	51	44	56	56	45	47	60	55	14
AV	69	62	26	53	75	55	56	60	60	87	52	64	66	79
None	56	56	43	46	52	28	45	44	44	11	42	42	37	46
Aver.(A,V,AV)	61	60	49	54	**69**	58	56	47	47	55	42	62	**64**	56
6 Movies	65	65	53	58	**68**	67	60	58	58	64	54	**65**	**65**	64

Table 1: F-scores $\left(F_{score}^{-1} = P_{recision}^{-1} + R_{ecall}^{-1} \right)$ for the HMM Variant and the Max-Product Dynamic System (MPDS) using either the GMM estimated or the bottom-up likelihoods. For the operation $\star$ we have employed three different versions: product, minimum and maximum.

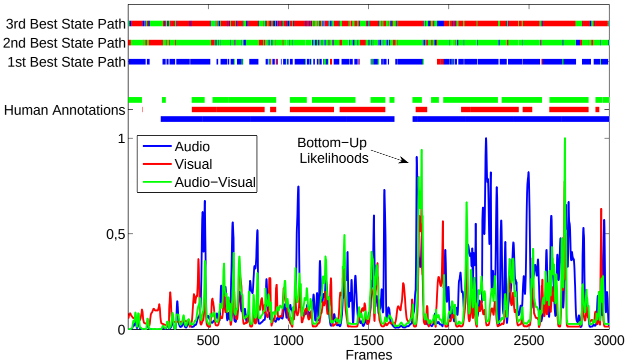

Fig. 1: Evolution of audio (blue), visual (red) and audio-visual (green) bottom-up likelihoods. We also see the human annotations and the 3-Best state paths using the Max-Product Dynamic System (MPDS) with product operation. (This figure is best viewed in color.)

5. CONCLUSIONS

We have developed a theory for max-product systems based on complete weighted lattices. Results of the theoretical analysis include analytic formulae for their state and output responses, max-product convolutions connecting inputs with outputs, and study of control-theoretic issues. Further, we have applied max-product systems to extend the Viterbi algorithm in HMMs to a more general scenario that allows for high-level control inputs in addition to the observations. This control-based new version of HMMs was applied to estimate audio-visual saliency states in multimodal videos. Comparisons between the results of the max-product system and human-annotations on movie videos yielded promising results for automatically detecting salient events. Our ongoing and future work in this area includes a further study of the relationship between the max-product dynamical systems and HMMs and development of approaches for estimating the max-product system parameters and state from observed data.

6. REFERENCES

[1] R. Bellman and W. Karush. On the maximum transform. *J. Math. Anal. Appl.*, 6:67–74, 1963.

[2] G. Birkhoff. *Lattice Theory*. Amer. Math. Soc., Providence, Rhode Island, 1967.

[3] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] R. W. Brockett. *Finite-dimensional Linear Systems*. Wiley, 1970.

[5] R. W. Brockett. Language driven hybrid systems. In *Proc. 33rd Conf. Decision & Control*, 1994.

[6] W. L. Brogan. *Modern Control Theory*. Quantum Publ., New York, 1974.

[7] P. Butkovič. *Max-linear Systems: Theory and Algorithms*. Springer, 2010.

[8] X.R. Cao and Y.C. Ho. Models of discrete event dynamical systems. *IEEE Control Syst. Magazine*, pages 69–76, Jun. 1990.

[9] C. G. Cassandras and S. Lafortune. *Introduction to Discrete Event Systems*. Kluwer Acad. Publ., 1999.

[10] G. Cohen, D. Dubois, J.P. Quadrat, and M. Viot. A linear system theoretic view of discrete event processes and its use for performance evaluation in manufacturing. *IEEE Trans. Automatic Control*, 30:210–220, 1985.

[11] G. Cohen, P. Moller, J.P. Quadrat, and M. Viot. Algebraic Tools for the Performance Evaluation of Discrete Event Systems. *Proc. IEEE*, 77:39–58, 1989.

[12] R. Cuninghame-Green. *Minimax Algebra*. Springer-Verlag, 1979.

[13] A. Doustmohammadi and E. W. Kamen. Direct generation of event-timing equations for generalized flow shop systems. In *Modeling, Simulation, and Control Technologies for Manufacturing*, volume 2596 of *Proc. SPIE*, pages 50–62, 1995.

[14] G. Evangelopoulos and P. Maragos. Multiband modulation energy tracking for noisy speech detection. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 14(6):2024–2038, Nov. 2006.

[15] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Trans. Multimedia*, 15(7):1553–1568, Nov. 2013.

[16] M. J. Gazarik and E. W. Kamen. Reachability and observability of linear systems over max-plus. *Kybernetika*, 35(1):2–12, 1999.

[17] S. Haykin, editor. *Kalman Filtering and Neural Networks*. J. Wiley & Sons, 2001.

[18] H.J.A.M. Heijmans. *Morphological Image Operators*. Acad. Press, Boston, 1994.

[19] H.J.A.M. Heijmans and C. Ronse. The algebraic basis of mathematical morphology. part i: Dilations and erosions. *Computer Vision, Graphics, and Image Processing*, 50:245–295, 1990.

[20] T. Hori and A. Nakamura. *Speech Recognition Algorithms Using-Weighted Finite-State Transducers*. Morgan & Claypool, 2013.

[21] V.G. Kaburlasos and V. Petridis. Fuzzy Lattice Neurocomputing (FLN) Models. *Neural Networks*, 13:1145–1169, 2000.

[22] T. Kailath. *Linear Systems*. Prentice-Hall, 1980.

[23] E. W. Kamen. An equation-based approach to the control of discrete even systems with applications to manufacturing. In *Proc. Int'l Conf. on Control Theory & its Applications*, Jerusalem, Israel, Oct. 1993.

[24] E. W. Kamen and A. Doustmohammadi. Modeling and stability of production lines based on arrival-to-departure delays. In *Proc. 33rd Conf. Decision & Control*, 1994.

[25] G. J. Klir and B. Yuan. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice-Hall, 1995.

[26] Y. Lucet. What shape is your conjugate? a survey of computational convex analysis and its applications. *SIAM Review*, 52(3):505–542, 2010.

[27] P. Maragos. Lattice image processing: A unification of morphological and fuzzy algebraic systems. *J. Math. Imaging and Vision*, 22:333–353, 2005.

[28] P. Maragos. Morphological filtering for image enhancement and feature detection. In A.C. Bovik, editor, *Image and Video Processing Handbook*, pages 135–156. Elsevier Acad. Press, 2 edition, 2005.

[29] P. Maragos. Representations for morphological image operators and analogies with linear operators. In P.W. Hawkes, editor, *Advances in Imaging and Electron Physics*, volume 177, pages 45–187. Acad. Press: Elsevier Inc., 2013.

[30] P. Maragos. Dynamical systems on weighted lattices. *IEEE Trans. Automatic Control*, submitted.

[31] P. Maragos, G. Stamou, and S. G. Tzafestas. A lattice control model of fuzzy dynamical systems in state-space. In J. Goutsias, L. Vincent, and D. Bloomberg, editors, *Mathematical Morphology and Its Application to Image and Signal Processing*. Kluwer Acad. Publ., Boston, 2000.

[32] M. Mohri, F. Pereira, and M. Ripley. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 16:69–88, 2002.

[33] H. Ney and S. Ortmanns. Progress in Dynamic Programming Search for LVCSR. *Proc. IEEE*, 88(8):1224–1240, Aug. 2000.

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

[35] R. T. Rockafellar. *Convex Analysis*. Princeton Univ. Press, Princeton, 1970.

[36] J. Serra. *Image Analysis and Mathematical Morphology*. Acad. Press, 1982.

[37] J. Serra, editor. *Image Analysis and Mathematical Morphology*, volume 2: Theoretical Advances. Acad. Press, 1988.

[38] S. R. Sternberg. Grayscale morphology. *Computer Vision, Graphics, and Image Processing*, 35:333–355, 1986.

[39] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vis.*, 57(2):137–154, 2004.

[40] Y. Weiss and W. T. Freeman. On the optimality of solutions of the max-product belief-propagation algorithm in arbitrary graphs. *IEEE Trans. Information Theory*, 35:736–744, 2001.