# Efficient Computation of the Frame-based Extend Union Model and its Application in Speech Recognition against Partial Temporal Corruptions

Yu-Chung Chan and Manhung Siu*
Hong Kong University of Science and Technology, EEE Department,
Clearwater Bay, Hong Kong.
Phone: +852 2358 8027, Fax: +852 2358 1485
Email: arthur@cs.ust.hk, msiu@ieee.org

**Abstract**

The extended union model (EUM) was proposed and shown to be effective in handling short time temporal corruption. Because of the computational complexity, the EUM probability can only be computed over groups of consecutive observations (called segments) and recognition can only be performed under N-best re-scoring paradigm. In this paper, we introduce a hidden variable called "pattern of corruption" and re-formulate the extended union model as marginalizing over possible patterns of corruption with likelihood computed via the missing feature theory. We then introduce a recursive relationship between the EUM probabilities of two successive observation sequences that can greatly simplify the EUM probability computation. This makes it possible to compute the EUM probability over a long sequence. Using this recursive relationship, the EUM probability over frames, called the "frame-based EUM" can easily be computed. To simplify the EUM-based recognition, we propose an approximated, dynamic programming-based EUM recognition algorithm, called the Frame-based EUM Viterbi algorithm (FEVA), that performs recognition directly instead of via N-best re-scoring. Experimental results on digit recognition under added impulsive noises show that both the frame-base EUM and the FEVA outperform the segment-based EUM.

## I. INTRODUCTION

As speech recognition matures and is being fielded in real world environments, robustness against unknown noisy environments becomes vital. Some real-life noises, such as door slams or telephone rings, are short and impulsive in nature. Their frequencies of occurrence and durations are difficult to predict.

Because these corrupted observations do not match well with the clean speech models, they would bias the observation likelihood and cause recognition errors. The problem of handling corrupted observations is recently addressed by missing feature theory [1], which proposed to omit the corrupted portion of the observation sequence from the likelihood computation. Other authors suggested that the corrupted components can be estimated from the clean components either in the spectral domain [2] or even in the general spectral-temporal domain [3]. It is well-known that prior knowledges such as location of corruptions and the number of corruptions can greatly improve the robustness of the algorithms [4].

The probabilistic union model, first introduced by Ming and Smith [5], is a general approach to reduce the model's sensitivity to unknown corruptions by using the probabilistic union operation. It was found to be very useful in dealing with partial corruptions. It was adapted to handle corruptions in sub-band-level [6] or word-level [7]. In [8], it was generalized into the extended union model (EUM) to make it applicable to robust speech recognition against *partial temporal corruptions*, or corruptions which affect only a subset of the observations This class of noise includes door-slam, keyboard click sounds and packet loss in digital transmission network.

The EUM is a probabilistic union of all possible observation subsets. For a $T$-frame observation sequence in which $M$ observations are corrupted by short time noise at unknown positions, there are $C_M^T$ possible subsets of $T-M$ observations. One of these subset contains no corrupted observation. Ideally, if the identities of the corrupted observations are known and the joint observation probability of the observation sequence is computed by excluding the corrupted observation as suggested in [1], then, this joint observation probability would be the same as the probability of the clean subset. In reality, the identities of the corrupted observations are often not known. The EUM probability of the $T$-frame observation sequence is defined as the probability of the union of all the $C_M^T$ possible subsets and can be computed as a sum over the probabilities of each subset. By assuming that the probability of the corrupted observations would be small compared to the clean observations, this sum would be dominated by the probability of the clean subset making the EUM probability a close approximation of ideal probability.

Using the EUM probability as a recognition criterion would require solving two computationally intensive problems. First, one needs to compute the EUM probability of a given state sequence. As will be discussed in Section II, because the EUM is combinatorial in nature, direct computation of the EUM probability of a long observation sequence (large $T$) is infeasible computationally. Second, in speech recognition, one needs to evaluate the EUM probabilities over all possible state sequences.

Two approximations were proposed in [8] to handle the computational issues. First, the observation sequence can be partitioned into a sequence of fixed-length segments [1]. Since the number of segments would be substantially smaller than the number of frames, direct computation of the EUM probability would be feasible. Second, recognition can be performed under the $N$-best re-scoring paradigm [9] to reduce the number of state sequences to be evaluated. By using these approximations, recognition using the EUM probability was reported to perform better than using the joint observation likelihood when the test utterances were corrupted by short-time impulsive noises. The segment length holds the balance between performance and computation. Using a long segment reduces computation. But, it could either ignore short corruptions or mark a segment corrupted even though only few observations (frames) in the segment are corrupted. Using short segments would increase modeling resolution but is computationally expensive. The choice of the segment length have to be optimized for different noise conditions empirically (Section

[1]Here, the term "segment" is not a representation of an acoustic unit as in segmental modeling. It is simply a sequence of consecutive frames. The usage of term follows [8].

4.2 [8]) which imposes some constraints on the general applicability of the approach. Long segments are typically used to reduce computation. To alleviate the side effect of long segments, [10] suggested shifting the segment boundaries such that the first segment would not always begin at the first observation.

Single-observation segments in the EUM (or the frame-based EUM) would be ideal if one can solve the computational problem. In this paper, we propose an efficient algorithm to compute the EUM probability that makes the frame-based EUM feasible. This algorithm centers on a recursive relationship between the EUM probability of a $t+1$-frame sequence and the EUM probability of the first $t$-frame sub-sequence. By means of direct computation, computation complexity in obtaining the EUM probability would be exponential on the sequence length $T$. By using this recursion, the complexity is reduced to a linear function of $T$.

To take full advantage of the EUM framework, it is desirable to use the EUM probability directly in Viterbi recognition instead of under the N-best re-scoring re-scoring paradigm. In this paper, we derive another recursive relationship between the EUM probabilities across different state sequences and devise an approximate dynamic-programming-based algorithm called the **f**rame-based **e**xtended union model **V**iterbi **a**lgorithm (FEVA) which is an approximation to the full EUM search. The FEVA provides an alternative solution to the two-stage recognition scheme described in [8].

This paper is organized as follows. In Section II, we review the EUM formulation and express it as a marginalization over different patterns of corruption. We also describe how the EUM can be used in speech recognition. In Section III, we describe the first recursive relation mentioned above and discuss how the computation of the EUM probability can be simplified. The FEVA will then be described in Section IV. In Section V, we evaluate the proposed algorithms under partial temporal corruptions. We summarize and conclude the paper in Section VI.

## II. The Extended Union Model for Speech Recognition

Denote a sequence of $T$ observations as $O_1^T = (o_1, \ldots, o_T)$. The conditional likelihood of this sequence for a given state sequence $Q_1^T$ can be expressed as

$$\begin{aligned} P(O_1^T|Q_1^T) &= p(o_1, \ldots, o_T|Q_1^T) \\ &= \prod_{t=1}^{T} p(o_t|q_t). \end{aligned}$$

Suppose out of the $T$ frames, $M$ are corrupted at unknown locations. Denote $l_i = \{t_{i,1}, \ldots, t_{i,M}\}$ as a possible set of $M$ observation time indexes such that $o_{t_{i,k}}$ is corrupted. For example, to mark $o_2, o_4, o_7$ as corrupted, $l = \{2, 4, 7\}$. By rules in combinatorics, there are $c_M^T$ possible $l_i$, i.e. $1 \le i \le c_M^T$. We call the $l_i$'s the **patterns of corruption**. If the pattern of corruption is known, then according to the missing feature theory [1], the corrupted observations should be ignored in the calculation of the joint likelihood. That is,

$$P(O_1^T|Q_1^T, l_i) \doteq \prod_{t=1, t \notin l_i}^{T} p(o_t|q_t).$$

This joint probability of only the known clean frames can be used as the optimization function in recognition.

In reality, $l_i$ is not known. Instead, *all possible* patterns of corruption can be considered by using their **union**. Denote the union of the observations across all possible corruption patterns as ${\mathring{O}_{\cup(T)}^{M}}^{2}$ in which the $M$ denotes the number of corrupted frames and $T$ denotes the total number of frames in the sequence.

$$\mathring{O}_{\cup(T)}^{M} = O_1^T \wedge (l_1 \cup l_i \ldots \cup l_{c_M^T}).$$

This is called an extended union model (EUM) of order $M$, which describes the union event of all possible patterns with $M$ corruptions.

Because the patterns of corruption are mutually exclusive, the $p(\mathring{O}_{\cup(T)}^{M})$ can be expressed as a sum of the probability of the observations and the individual patterns. That is,

$$\begin{aligned} p(\mathring{O}_{\cup(T)}^{M}|Q_1^T) &= p(O_1^T \wedge (l_1 \cup l_i \ldots \cup l_{c_M^T})|Q_1^T) \\ &= \sum_{i=1}^{c_M^T} p(O_1^T, l_i|Q_1^T) \\ &= \sum_{i=1}^{c_M^T} p(O_1^T|l_i, Q_1^T)p(l_i|Q_1^T) \\ &= \sum_{i=1}^{c_M^T} p(O_1^T|l_i, Q_1^T)p(l_i) \\ &= \frac{1}{c_M^T} \sum_{i=1}^{c_M^T} p(O_1^T|l_i, Q_1^T). \end{aligned} \tag{1}$$

The above formulation assumes that the pattern of corruption is not dependent on the state sequence which is equivalent to assuming that the noise is independent of the speech signal. Furthermore, with no

[2]Strictly speaking, because the EUM probability is calculated from frame 1 to frame $t$, a more vigorous notation would be $(\mathring{O}_1^t)_{\cup(t)}^{M}$. Since we typically assumes that the beginning frame is 1 and that the ending index is always the same as the total number of frames considered, we adopt a simpler notation $\mathring{O}_{\cup(t)}^{M}$ in this paper.

| $M$ | $p(\mathring{O}^{M}_{\cup(T)})$ |
|---|---|
| 0 | $p(o_1 o_2 o_3 o_4)$ |
| 1 | $p(o_1 o_2 o_3 \cup o_1 o_2 o_4 \cup o_1 o_3 o_4 \cup o_2 o_3 o_4)$ |
| 2 | $p(o_1 o_2 \cup o_1 o_3 \cup o_1 o_4 \cup o_2 o_3 \cup o_2 o_4 \cup o_3 o_4)$ |
| 3 | $p(o_1 \cup o_2 \cup o_3 \cup o_4)$ |

TABLE I

EXAMPLE OF THE EUM PROBABILITY OF DIFFERENT ORDERS IN THE CASE FOR 4 FRAMES, ORDER $0, 1, 2, 3$ REPRESENT $0, 1, 2, 3$ FRAMES ARE MISSING. NOTE THAT THE STATE INFORMATION WAS IGNORED HERE.

prior knowledge of the noise, the prior probability of the pattern of corruption, $p(l_i)$, is assumed to be uniform.

Why would the EUM probability be robust to corruptions? One can view this from two perspectives. First, corrupted observations do not match the model well and would result in lower likelihood than the clean observations. This implies that $P(\mathring{O}^{M}_{\cup(T)}|Q_1^T)$ would be dominated by the term $p(O_1^T|l_{\hat{i}}, Q_1^T)$ which contains well matched clean observations. Then, the effect of the corruptions is reduced in the decision process. The idea is closely related to robust statistics: If we regard the process of likelihood computation as gathering statistics from individual frames, Equation 1 can be interpreted as jackknife estimates [11] of likelihood that leaves $M$ observations out.

The second perspective is suggested in [8]. They observed that the likelihood function, which involved an "and" operation on the observations which is good for determining the correctness of all the frames against a model but is sensitive to corruption of individual frames. A probabilistic general union model, on the other hand, involves an "or" operation on observations that would average out the observation probabilities of corrupted frames and clean frames. The EUM works as a trade-off between the above the two cases. As $M$ becomes smaller, the EUM approaches the joint likelihood function. As $M$ becomes larger, the EUM approaches a general union model. By properly adjusting the value of $M$, the EUM should be able to cope with any pattern of corruption.

As we can see in Equation 1, the computation of $P(\mathring{O}^{M}_{\cup(T)}|Q_1^T)$ involves $c_M^T$ terms. Table 1 shows the EUM of order $0, 1, 2, 3$ for a four-frame sequence. (We ignore the state information in the table and only show the observation information there.) We can see that direct implementation can be computationally intractable for large $T$ ($T > 100$) which is the typical length of a speech utterance. In Section III, we describe how this can be solved more efficiently.

In a HMM-based speech recognition system, instead of considering only one state sequence, we need to find the most likely state sequence, $\hat{Q}_1^T$, given the observations $O_1^T$, i.e.,

$$\hat{Q}_1^T = \arg\max_{Q_1^T} p(Q_1^T | O_1^T). \quad (2)$$

By using Bayes rule and using the fact that $P(O_1^T)$ is independent of the state sequence, Equation 2 can be re-written as maximizing the joint observation and state likelihood, i.e.,

$$\begin{aligned} \hat{Q}_1^T &= \arg\max_{Q_1^T} p(O_1^T, Q_1^T) \\ &= \arg\max_{Q_1^T} p(O_1^T | Q_1^T) p(Q_1^T). \end{aligned} \quad (3)$$

Under the EUM framework, the observation likelihood is replaced by the EUM probability of order $M$ that is robust to $M$ corruptions. I.e.

$$\hat{Q}_1^T = \arg\max_{Q_1^T} p(\mathring{O}_{\cup(T)}^M | Q_1^T) p(Q_1^T). \quad (4)$$

Equation 4 poses another computation difficulty: The need to search for the state sequence with highest EUM probability. While the Viterbi algorithm can efficiently find the best state sequence to solve Equation 3, it is not clear how to search for the best state sequence that optimized the EUM probability in Equation 4.

To reduce the number of state sequences to be searched, [8] suggested applying the N-best re-scoring paradigm. This simplifies the problem of evaluating the EUM probabilities of all state sequences into evaluating $N$ hypotheses only at the expense of a more complex decoding process and a dependence on the quality of the N-best list. Furthermore, to reduce $T$ in the EUM, the observations sequence was partitioned into fixed-length segments of length $D$, denoted as $(z_1, \ldots, z_{\lceil \frac{T}{D} \rceil})$. Because the number of segments, $\lceil \frac{T}{D} \rceil$ is significantly smaller than $T$, the best state sequence $\hat{Q}_1^T$, which optimized the segment-based extended union probability with $M$ segments being corrupted can be found. The segment based EUM recognition equation is given by

$$\hat{Q}_1^T = \arg \max_{Q_1^T \in (\text{N-best})} P(\mathring{Z}_{\cup(M)}^{\lceil \frac{T}{D} \rceil}, Q_1^T) \quad (5)$$

Partitioning the observations to segments may have some drawbacks. If part of the segment was corrupted, the whole segment could either be regarded as corrupted or completely clean. This could result in either excess data loss or reduced robustness. Another drawback is that setting the segment length can be difficult. To reduce the side effect of long segment, [10] suggested the use of a “variable” segmentation that shifts the beginning of the segments while keeping the segment length fixed. In general, the EUM over individual frames is desirable but is hindered by its computation complexity. We will discuss how this problem can be solved in the next section.

## III. Fast Computation of the EUM Likelihood

As we have mentioned in the previous section, computing the EUM probability of a given state sequence using Equation 1 directly is computationally expensive for any moderate $T$. However, many of the computations are repeated. Similar to the dynamic programming approach applied in Viterbi and forward-backward algorithm, a recursive relationship can be established between the EUM probability of $t$ observations and $t-1$ observations. By creating such a recursion, the computation complexity can be reduced from exponential to a linear order.

Using the definition of the EUM in Section II Equation 1, the EUM probabilities of a $t$-frame sequence with $m$ corruptions given a state sequence $Q_1^t$ can be expressed as

$$p(\mathring{O}^M_{\cup(t)}|Q_1^t) \quad = \quad \frac{1}{c_m^t}\sum_{i=1}^{c_m^t} p(O_1^t|l_i, Q_1^t), \tag{6}$$

where $l_i$ is the $i$-th pattern of corruption. Within the $c_m^t$ patterns of corruption, some patterns mark observation $o_t$ as corrupted and some do not. Thus, we can divide the $c_m^t$ patterns into two subsets, $L_1$ and $L_2$. The first set, $L_1$, includes only those patterns with $o_t$ marked as corrupted and ignored in the joint probability. The second set, $L_2$, is the complement of $L_1$.

Because the sets $L_1$ and $L_2$ are mutually exclusive, Equation 6 can be rewritten as

$$P(\mathring{O}^m_{\cup(t)}|Q_1^t) \quad = \quad \frac{1}{c_m^t}\left(\sum_{l_i\in L_1} p(O_1^t|l_i, Q_1^t) + \sum_{l_i\in L_2} p(O_1^t|l_i, Q_1^t)\right). \tag{7}$$

For $L_1$, since $o_t$ is marked as corrupted (and ignored in likelihood computation), there are only $m-1$ corruptions out of the previous $t-1$ observation. Thus,

$$\begin{aligned} \sum_{l_i\in L_1} p(O_1^t|l_i, Q_1^t) \quad &= \quad \sum_{i=1}^{c_{m-1}^{t-1}} p(O_1^{t-1}|l_i, Q_1^t) \\ &= \quad c_{m-1}^{t-1} P(\mathring{O}^{m-1}_{\cup(t-1)}|Q_1^{t-1}). \end{aligned} \tag{8}$$

Notice that the conditioning on $Q_1^t$ is changed to $Q_1^{t-1}$ because the knowledge of $q_t$ is not needed when $o_t$ is marked as corrupted and being ignored in likelihood computation.

The corruption patterns in $L_2$ do not mark $o_t$ as corrupted. This means that there are $m$ corruptions out of the first $t-1$ observations. Also, observation $o_t$ is included in the likelihood computation. Thus,

$$\sum_{l_i \in L_2} p(O_1^t|l_i, Q_1^t) = \sum_{i=1}^{c_m^{t-1}} p(o_t|l_i, Q_1^t) p(O_1^{t-1}|l_i, Q_1^t) \tag{9}$$

$$= p(o_t|Q_1^t) \sum_{i=1}^{c_m^{t-1}} p(O_1^{t-1}|l_i, Q_1^t) \tag{10}$$

$$= p(o_t|q_t) \sum_{i=1}^{c_m^{t-1}} p(O_1^{t-1}|l_i, Q_1^{t-1}) \tag{11}$$

$$= c_m^{t-1} p(o_t|q_t) P(\mathring{O}^m_{\cup(t-1)}|Q_1^{t-1}). \tag{12}$$

Because of conditional independence assumption, the likelihood of the current observation $o_t$ can be decomposed as expressed in Equation 9. By definition, all the corruption patterns in $L_2$ do not include $t$, implying that the likelihood of $o_t$ is always included in the sum and is independent of the corruption patterns. Thus, Equation 9 can be rewritten into Equation 10. By using the conditional independence assumption again, $p(o_t|Q_1^t)$ is simplified into $p(o_t|q_t)$ and $p(O_1^{t-1}|l_i, Q_1^t)$ into $p(O_1^{t-1}|l_i, Q_1^{t-1})$ in Equation 11. Finally, using Equation 6, the sum in Equation 11 is converted into the EUM probability at time $t-1$ in Equation 12. Combining Equations 7-12 and using the fact that

$$\frac{c_{m-1}^{t-1}}{c_m^t} = \frac{m}{t}$$

and

$$\frac{c_m^{t-1}}{c_m^t} = \frac{t-m}{t},$$

we have the following recursion.

$$\begin{aligned} P(\mathring{O}^m_{\cup(t)}|Q_1^t) &= \left(\frac{c_{m-1}^{t-1}}{c_m^t}\right) P(\mathring{O}^{m-1}_{\cup(t-1)}|Q_1^{t-1}) + \left(\frac{c_m^{t-1}}{c_m^t}\right) P(\mathring{O}^m_{\cup(t-1)}|Q_1^{t-1}) p(o_t|q_t) \\ &= \left(\frac{m}{t}\right) P(\mathring{O}^{m-1}_{\cup(t-1)}|Q_1^{t-1}) + \left(\frac{t-m}{t}\right) P(\mathring{O}^m_{\cup(t-1)}|Q_1^{t-1}) p(o_t|q_t). \end{aligned} \tag{13}$$

We call Equation 13 the **EUM formula**. It has two recursive variables, $m$ and $t$. To compute $P(\mathring{O}^m_{\cup(t)}|Q_1^t)$, one would need $P(\mathring{O}^{m-1}_{\cup(t-1)}|Q_1^{t-1})$ at time $t-1$ which in turn depends on $P(\mathring{O}^{m-1}_{\cup(t-2)}|Q_1^{t-2})$. In general, using the EUM formula for computing $P(\mathring{O}^M_{\cup(T)}|Q_1^T)$, we need to compute $P(\mathring{O}^m_{\cup(t)}|Q_1^t)$ for time $t \leq T$ and all order $m \leq M$. Each update of $P(\mathring{O}^m_{\cup(t)}|Q_1^t)$ requires three multiplications and one addition. The computation, therefore, is of $O$(MT) instead of $O(C_T^M)$ as would be the case in the direct computation using Equation 1.

The same principle as in Equation 13 can be applied to the segment-based EUM as described in [8]. In addition to being useful for the segment-based EUM, the EUM formula can potentially be useful for other EUM applications as well.

## IV. The Frame-based EUM Viterbi Algorithm (FEVA)

The proposed recursion in Section III simplifies the computation of the EUM probability of a given state sequence and thus removes the constraint on segment length. This makes the frame-based EUM possible. However, in order to use it for speech recognition, N-best re-scoring paradigm that reduces the number of state sequence is still needed. While the N-best re-scoring paradigm has been used extensively for testing out new models, its usefulness is dependent on the quality of the n-best and a long n-best list actually can be expensive to re-score. Insights from other works such as [12] on robust speech recognition suggested that it is possible to devise a dynamic algorithm which directly optimizes the EUM probability.

For an HMM-based recognition system, efficient solution for Equation 2 can be obtained using the Viterbi algorithm by computing the partial path score $\delta_t(j)$, which is the best likelihood of $O_1^t$ and state sequence up to time $t-1$ such that $q_t = j$. Mathematically,

$$\delta_t(j) \quad = \quad \max_{Q_1^{t-1}} P(O_1^t, Q_1^{t-1}, q_t = j) \tag{14}$$

A recursive relation on $\delta_t(j)$ in the Viterbi algorithm is given by

$$\delta_t(j) \quad = \quad \max_i \left[\delta_{t-1}(i) a_{ij}\right] b_j(o_t) \quad 1 \leq t \leq T, \tag{15}$$

where $b_j(o_t)$ is the observation probability of $o_t$ at state $j$.

In Equation 13, we have a recursive relationship across time for a given state sequence. In order to make the EUM-based recognition possible, a recursion across different state sequences similar to Equation 15 would be needed.

### A. Computing the EUM Probability

We define the partial path score $\delta_t(j, m)$ as the EUM likelihood at time $t$ for state $j$ with the best state sequence up to time $t-1$ for an order $m$ EUM. Mathematically, $\delta_t(j, m)$ can be expressed as

$$\delta_t(j, m) \quad = \quad \max_{Q_1^{t-1}} p(\mathring{O}_{\cup(t)}^m, Q_1^{t-1}, q_t = j) \tag{16}$$

If we denote $\nu_m(t) = \frac{m}{t}$, then, an approximated algorithm is derived as follows.

$$
\begin{aligned}
&\delta_t(j,m) \\
&= \max_{Q_1^{t-1}} p(\mathring{O}^m_{\cup(t)}, Q_1^{t-1}, q_t = j) \\
&= \max_i \left[ \max_{Q_1^{t-2}} \left( p(\mathring{O}^m_{\cup(t)}, Q_1^{t-2}, q_{t-1} = i, q_t = j) \right) \right] \\
&= \max_i \left[ \max_{Q_1^{t-2}} \left( p(\mathring{O}^m_{\cup(t)} | Q_1^{t-2}, q_{t-1} = i, q_t = j) p(Q_1^{t-2}, q_{t-1} = i, q_t = j) \right) \right] \\
&= \max_i a_{ij} \left[ \max_{Q_1^{t-2}} \left( p(Q_1^{t-2}, q_{t-1} = i) p(\mathring{O}^m_{\cup(t)} | Q_1^{t-2}, q_{t-1} = i, q_t = j) \right) \right] \qquad (17) \\
&= \max_i a_{ij} \left[ \max_{Q_1^{t-2}} \left( p(Q_1^{t-2}, q_{t-1} = i) \right.\right. \\
&\qquad \left.\left. \left\{ \nu_m(t) p(\mathring{O}^{m-1}_{\cup(t-1)} | Q_1^{t-2}, q_{t-1} = i) + (1-\nu_m(t)) p(\mathring{O}^m_{\cup(t-1)} | Q_1^{t-2}, q_{t-1} = i) b_j(o_t) \right\} \right) \right] \qquad (18) \\
&= \max_i a_{ij} \left[ \max_{Q_1^{t-2}} \left( \nu_m(t) p(\mathring{O}^{m-1}_{\cup(t-1)}, Q_1^{t-2}, q_{t-1} = i) + (1-\nu_m(t)) p(\mathring{O}^m_{\cup(t-1)}, Q_1^{t-2}, q_{t-1} = i) b_j(o_t) \right) \right] \qquad (19) \\
&\approx \max_i a_{ij} \left[ \nu_m(t) \max_{Q_1^{t-2}} p(\mathring{O}^{m-1}_{\cup(t-1)}, Q_1^{t-2}, q_{t-1} = i) \right. \\
&\qquad \left. + (1-\nu_m(t)) \max_{Q_1^{t-2}} p(\mathring{O}^m_{\cup(t-1)}, Q_1^{t-2}, q_{t-1} = i) b_j(o_t) \right] \qquad (20) \\
&= \max_i a_{ij} \left[ (\nu_m(t) \delta_{t-1}(i, m-1) + (1-\nu_m(t)) \delta_{t-1}(i,m) b_j(o_t)) \right] \qquad (21)
\end{aligned}
$$

By means of Bayes rule and the Markov assumption, we can rewrite Equation 16 into Equation 17 such that $\delta_t(j,m)$ depends on quantities at time $t-1$. Equation 18 is obtained by applying the recursion from Equation 13 such that the EUM probability at time $t$ can be expressed as a sum of the EUM probabilities at time $t-1$. By applying the Bayes rule Equation 18 can be rewritten into Equation 19 which is a maximization over a sum of two terms at time $t-1$. Notice that $\delta_t(j,m)$ is now expressed as the score of a **single** state sequence $Q_1^{t-2}$ that maximizes the sum of the two terms. However, each term by itself is not necessarily the best score at time $t-1$ at state $i$. This implies that to compute $\delta_t(j,m)$ exactly, one has to store the partial path scores for all state sequence $Q_1^{t-2}$ rendering the recursion useless. The maximum of the sum, however, can be approximated by the sum of the maximum in Equation 20. This in effect is an optimistic estimation of the EUM path score. Then, a recursion on $\delta_t(j,m)$ can be obtained in Equation 21.

There are two recursion variables $m$ and $i$ in Equation 21. $\delta_t(j,m)$ depends on the $\delta$'s at $t-1$ from different states $i$ of orders $m$ and $m-1$. One can view the above recursion as expanding the path metric $\delta_j$ into a $M+1$ dimensional vector. Hence, the recursion can be simply implemented by expanding the search space $M+1$ times and applying Equation 21. We called this algorithm the frame-base EUM Viterbi algorithm (FEVA).

## *B. Back-tracking*

In standard Viterbi algorithm, determination of the best state sequence requires only state information in the previous time $t-1$. In the FEVA, back-tracking requires both the best previous state and the knowledge of the number of previous order $M_{t-1}$ because of the two recursion variables in Equation 21. Define $\psi_t(j,m)$ as the best previous state for an order $m$ EUM at time $t$ and state $j$.

$$\psi_t(j,m) = \arg\max_i a_{ij}[(1-\nu_m(t))\delta_i(t-1,m)b_j(o_t) + \nu_m(t)\delta_i(t-1,m-1)], \tag{22}$$

where $\nu_m(t) = \frac{m}{t}$. However, because of the approximation in Equations 20 and Equation 21, determination of the exact EUM order information is very difficult. This is because the $m$-th order partial score at time $t$ would depends on the sum of the $m$-th and $m-1$-th order EUM scores at time $t-1$ that can have different state histories. This situation is analogous to the difference between the forward-backward algorithm and the Viterbi algorithm. To search for an approximate best EUM state sequence, we propose the following approximation

$$M_t(m) \;=\; \arg\max_{m,m-1} \left\{(1-\nu_m(t))\delta_{\hat{i}}(t-1,m)b_j(o_t), \nu_m(t)\delta_{\hat{i}}(t-1,m-1)\right\}, \tag{23}$$

where $\hat{i} = \psi_t(j,m)$. This approximation would mean that the score computed is not that of the recognized path. This type of approach, however is also used in other recognition system. For example, [13] used the forward probability as the partial path metric while the most likely state was used in back-tracking.

## *C. Implementation of the Frame-based EUM Viterbi algorithm*

Similar to the N-best recognition algorithm proposed in [9], one can directly implement Equations 21-23 by augmenting the state to store $M+1$ scores at each time instance. Figure 1 shows the implementation of the FEVA in which each box represents an augmented state that stores the partial path score $\delta_j(t,m)$, $0 \le m \le M$. The original Viterbi path, or the path with the EUM order 0 is also stored as $m=0$. During each update, the scores from all possible incoming arcs are calculated using Equation 21. Similar

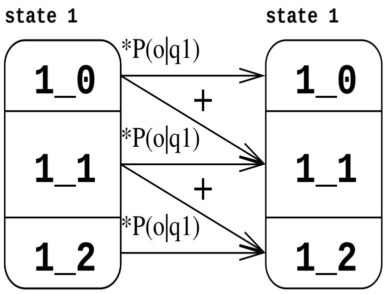


Fig. 1. Update diagram for the FEVA in a single-state HMM

to the Viterbi algorithm, using Equation 23, backtracking pointers can be used to record the state with maximum partial scores for each state. The dynamic range between $\nu_m(t)$ and the EUM probability of the previous frame is large. To simplify implementation, the contribution of $\nu_m(t)$ can be ignored.

It is possible that Equation 21 requires computation of union probabilities with either $m < 0$ or $t < 0$ or $m > t$. In all these cases, we can take $P(\mathring{O}^m_{\cup(t)}) = 0$.

The computation and memory size of this implementation would be $M$ times larger than a standard Viterbi algorithm. With the use of caching on the observation likelihood, the computation could be reduced to a similar level to the Viterbi algorithm.

## V. Experimental Results

In this section, we describe the experimental results of both the frame-based EUM and the FEVA. Similar to [8], experiments were conducted to evaluate recognition performance under partial temporal corruptions (PTC). A description of the corpus and experimental setup is given in Section V-A. Using N-best re-scoring paradigm, we compare recognition performance of the segment-based EUM, the frame-based EUM and the FEVA under wide ranges of noisy conditions in Section V-B.

### *A. Corpus and Conditions*

Before we describe the experimental conditions in detail, three terms, “frame”, “segment” and “block” are used in this section that can be confusing. “Frame” denotes an acoustic feature vector (MFCC). “Segment” denotes a sequence of consecutive frames under the EUM and is used in the context of the segment-based EUM. Finally, “block” denotes a short interval of noise that is added to speech to simulate the noisy environment. By using the above definitions, we hope to avoid ambiguity.

The experiments were carried out using the TI-digit [14] on the task of connected digit recognition. 8668 training utterances and 8668 testing utterances were used. Each utterance is a digit string with a maximum length of 7 digits and up to 3 seconds long. Speech observations (MFCC frames) were generated at a rate of 100 frames per second using a 25ms window. 39 dimensional front-end features wer used including 12 MFCC, frame energy, and their first and second order derivatives.

Each digit was represented by an 18-state left-to-right HMM. Silence was represented as a 5-state left-to-right HMM. Short pause was modeled using a 3-state t-model as described in [15]. We followed the training recipe described in [16] to estimate the parameters of the HMMs and obtained baseline word accuracy of 98.83% in connected digit recognition. [3]

We tested the algorithm in two conditions: clean and noise-corrupted. The original testing utterances in the TIDIGITS corpus were used in the clean condition. For the noise-corrupted cases, partially temporal corruptions (PTC) discussed in [8] were added. In our experiments, noises blocks were extracted from 4 telephone ring tones, 1 whistling noise and two message sound clips in ICQ. These noise blocks were randomly added to the testing utterances. To test the effectiveness of the proposed algorithms, we varied the following attributes of the noise corruptions.

- **Signal-to-noise ratio** (**$S$**) is computed as the ratio between the power of the speech signals and the noise added. In our experiments, the signal-to-noise ratio was computed based the following equation,
$$SNR = 10\log_{10}\left(\frac{\text{power of the speech signal}}{\text{power of the added noise}}\right), \tag{24}$$
with silences and short-pauses excluded in the calculation of the power of the speech signal.
- **Rate of corruption** (**$C$**) is the probability that a block of speech would be corrupted by additive noise.
- **Block Length** (**$L$**) is the duration of the block measured in mini-seconds (ms). In our experiments, block lengths are always multiples of 25ms.

In our experiments, each speech utterance was partitioned into non-overlapping blocks. For each block, a Bernoulli random variable with $p = C$ was generated to decide whether noise should be added. This simulates the general scenario in which short-time noise occurs in a time-varying and random manner. Because the noise is added in time domain, the use of overlapping window in cepstral generation and derivative computation results in corruption of multiple frames.

[3]The results was untuned, by tuning different penalties, higher accuracy above 99.05% can be reached.

### B. Digit Recognition Under Partial Temporal Corruptions

We first tested the frame-based EUM in the N-best re-scoring framework. 50 N-best hypotheses were generated using HTK [15]. The segment-based EUM and the frame-based EUM re-scored these 50-best utterances separately.

Because the length of each utterance varies in connected digit recognition, we followed the convention of [8] to use the notation of relative order, which is the ratio of the order of the EUM against the length of an utterance. For example, a relative order of 0.2 would mean that 20% of the frames (or segment) in an utterance is assumed to be corrupted. Similar to [8], we chose a relative order of 0.1 or 10% for all the experiments.

| Conditions | BL | Segment | Frame | FEVA |
|---|---|---|---|---|
| clean | 98.83 | 98.50 | 98.58 | 98.44 |
| -10dB,C=20%,L=25ms | 84.61 | 88.49 | 89.90 | 90.70 |

TABLE II

COMPARISON OF DIFFERENT ALGORITHMS, FRAME-BASED EUM, SEGMENT-BASED EUM, FEVA AND BASELINE VITERBI ALGORITHM, IN CLEAN AND NOISY CONDITIONS.

The frame-based and segmented-based EUMs were first compared in two conditions, clean and noisy with PTC added at $S = -10dB, C = 20\%, L = 25ms$. The results are summarized in Table II. For the segment-based experiments, a segment length of 6 frames were used. In the frame-based experiments, the segment length was effectively 1 frame. The frame-based EUM was possible computationally because of the efficient recursive formula described in Section IV. Under the noisy environments, both the segment-based and the frame-based EUM outperformed the Viterbi algorithm. Furthermore, the frame-based approach outperformed the segment-based approach in noisy condition by $12\%$ showing that a finer resolution of segment is beneficial for the EUM. In clean condition in which no corruption was present, both approaches hurt performance with similar amounts of degradation. This is consistent with the reported results in [8].

In the second set of experiments, the frame-based and segmented-based EUMs were tested under different partial temporal corruption conditions. Columns 3 and 4 in Tables III-V show the comparison under various SNRs, rates of corruption and lengths of corruption.

In Table III, the SNR was varied between $-10dB$ to $10dB$. Lower SNR caused more degradation but

| SNR | Viterbi | Seg-based EUM | Frame-based EUM |
|---|---|---|---|
| $\infty$ | 98.83 | 98.50 | 98.58 |
| 10 | 98.83 | 98.47 | 98.53 |
| 0 | 97.67 | 97.47 | 97.77 |
| -10 | 84.61 | 88.49 | 89.90 |

TABLE III

PERFORMANCES OF DIFFERENT ALGORITHMS IN PTC WITH VARYING SNR AND $C = 20\%, L = 25ms$ AT $\frac{M}{N} = 0.1$.

| Rate of Corruption | Viterbi | Seg-based EUM | Frame-based EUM |
|---|---|---|---|
| 20% | 84.61 | 88.49 | 89.90 |
| 30% | 68.92 | 74.28 | 76.82 |
| 40% | 59.06 | 64.17 | 68.44 |

TABLE IV

PERFORMANCES OF DIFFERENT ALGORITHMS IN PTC WITH VARYING RATES OF CORRUPTION AND $SNR = -10dB, L = 25ms$ WITH $\frac{M}{N} = 0.1$

also increased the usefulness of the EUM. For the segment-based EUM, it was better than the baseline Viterbi only when the SNR fell to -10dB. However, the frame-based EUM was useful even at 0 dB. In Table IV, the rate of corruption was varied between 20% to 40% at $SNR = -10dB$ and the corruption block length set at $L = 25ms$. As the rate of corruption increased, the ASR performance degraded but increased the effectiveness of the frame-based EUM. This effect is similar to what was observed with changing SNR. Furthermore, consistent with results in Table III, the frame-based EUM is always better than the segment-based EUM across the different rates of corruption. The results for varying the length of corruption from 25ms to 200ms are shown in Table V in which we kept $C = 20\%$ and $SNR = -10dB$. Shorter length of corruption caused more damage to the utterances if the rate of corruption was kept constant because more phonetic units would be affected. It is interesting to note that in the last experiment, as the assumption of long segment length became more accurate when the segment-length was longer than 100ms, the performances of the frame-based and the segment-based EUMs became more similar. However, the frame-based EUM was still slightly better the segment-based EUM in these cases. One

| length of corruption | Viterbi | Seg-based EUM | Frame-based EUM |
|---|---|---|---|
| 25ms | 84.61 | 88.49 | 89.90 |
| 50ms | 87.86 | 90.38 | 91.31 |
| 75ms | 90.82 | 92.88 | 93.72 |
| 100ms | 93.16 | 94.92 | 95.55 |
| 125ms | 95.31 | 96.55 | 96.69 |
| 150ms | 95.48 | 96.65 | 96.86 |
| 175ms | 95.35 | 96.47 | 96.59 |
| 200ms | 95.22 | 96.17 | 96.38 |

TABLE V

PERFORMANCES OF DIFFERENT ALGORITHMS WITH VARYING CORRUPTION DURATIONS AND $SNR = -10dB, C = 20\%$ WITH $M/N = 0.1$.

reason may be that while the added noise block is long and consistent with the segment assumption, it may fall between two segments thus, corrupting parts of two consecutive segments.

Comparing the EUM and the Viterbi algorithm, results shown in Tables III-V suggest that the EUM model is more powerful when the condition is worse (low SNR or high rate of corruption). Comparing the frame-base EUM and the segment-based EUM with a 6-frame segment, the frame-based EUM is consistently better than the segment-based EUM because of its finer resolution.

| Viterbi | Seg-based EUM | Frame-based EUM | FEVA |
|---|---|---|---|
| 92.23 | 94.06 | 94.63 | 94.70 |

TABLE VI

AVERAGE PERFORMANCE OF DIFFERENT ALGORITHMS IN PTC ACROSS DIFFERENT BLOCK LENGTH OF 25-200MS AT $SNR = -10dB$ WITH $\frac{M}{N} = 0.1$

Our next set of experiments focused on the evaluation of the FEVA. Table II also shows the performance of using the FEVA under clean to -10 dB noisy conditions. Compared to the frame-based EUM with N-best re-scoring, the FEVA gave comparable performance with better performance at -10dB and slightly worse performance under clean environment. However, the FEVA was better than the segment-based EUM.

Results from Table III-IV suggest that the most interesting case for using the EUM with acceptable recognition performance is at $SNR = -10$dB and rate $C = 20\%$. Similar to the experiments reported in Table V, we varied the noise block length at this setting. To make comparison easier, we tabulate the average performance across different block lengths in Table VI. The results show that the performance of the FEVA was comparable but no better than N-best re-scoring on the frame-based EUM probably because of the approximations taken. However, the FEVA outperformed the segment-based EUM and allowed the use of EUM probability directly in recognition without the need for an extra N-best generation pass. Furthermore, its performance is not dependent on the quality of n-best which can be badly degraded under difficult noisy environments.

## VI. CONCLUSIONS AND FUTURE WORK

Union model is one of the emerging acoustic modeling techniques that move beyond traditional HMM and has been shown to be effective in handling corrupted observations. The application of union model to speech recognition is hindered by two computational problems, namely, the computation of the EUM probability for a given state sequence and the search for the optimal EUM probability state sequence.

This paper analyses the properties of the EUM probability and addresses these two computation problems by proposing two recursive algorithms that avoid repeated computations.

The first algorithm is based on the EUM formula which exploits the recursive relationship between the EUM probabilities of a length $t$ sequence and a length $t+1$ sequence. The proposed algorithm reduces the computation complexity of the EUM probability from an exponential order on sequence length to a linear order, making it possible to compute the EUM probabilities over long sequences. Because of this, it is no longer needed to artificially divide a speech utterance into segments. Instead, the EUM probability over individual speech frames, which we call the frame-based EUM, is feasible. The frame-based EUM not only removes the need to select the segment length which can be difficult and dependent on the noise environments, it also outperforms the segment-based EUM experimentally over different impulsive noise conditions. It is also important to note that even at long durations of impulsive noise close to the segment-length, the frame-based EUM is consistently better than the segment-based EUM.

To solve the EUM recognition problem, this paper describes an approximate Viterbi-like algorithm, called the frame-based EUM Viterbi algorithm (FEVA). The FEVA can efficiently perform recognition using the EUM probability as the criterion. Experimentally, the FEVA performed as well as the N-best rescoring without the need for N-best rescoring which can add complexity to the recognition paradigm. Results from N-best re-scoring could also be sensitive to the quality of the N-best generation. The derivation of the FEVA is also valuable for further development of the EUM.

The results we presented in the experiment section assumed a fixed relative order. It would be more desirable to have ways to select the order of the EUM. Work in [17] made used of duration model to re-score results with different orders. In [12], the problem was treated as a model-selection problem. Similar approaches may be applicable to the frame-based EUM and would be one focus of our future study.

## VII. Acknowledgements

This work is partially supported by the Hong Kong Research Grant Council, CAG project number: CAG02/03.eg05. The views in this article are that of the authors and do not reflect the view of the sponsor.

## References


[1] R. Lippmann and B. Carlson, "Robust speech recognition with time-varying filtering interruptions and noise," in *Proceeding of the ASR workshop*, 1997.

[2] M. Cooke, A. Morris, and P. Green, "Missing data techniques for robust speech recognition," in *ICASSP*, 1997.

[3] B. Raj, R. Singh, and R. M. Stern, "Inference of missing spectrographic for robust speech recognition," in *ICSLP*, 1998.

[4] D. V. Johan, F. D. Wet, B. Cranen, and L. Boves, "Missing data techniques for robust speech recognition," in *Proceeding. Workshop on Robust Methods in Adverse Conditions*, Tampere, Finland, 1999.

[5] J. Ming and F. J. Smith, "Union: A new approach for combining sub-band observations for noisy speech recognition," in *IEEE Workshop on Robust Speech*, Finland, 1999.

[6] J. Ming, P. Jancovic, and F. Smith, "Robust speech recognition using probabilistic union models," *IEEE transaction on Speech and Audio Processing*, vol. 10, September 2002.

[7] E. Sicilia-Garcia, J. Ming, and F. J. Smith, "Individual word language models and the frequency approach," in *ICSLP*, September 2002.

[8] J. Ming and F. J. Smith, "Union: a model for partial temporal corruption of speech," *Computer Speech and Language*, vol. 15, 2001.

[9] R. Schwartz and Y. Chow, "The n-best algorithm: an efficient and exact procedure for finding n most likely sentence hypotheses," in *ICASSP*, Albuquerque, April 1990.

[10] J. Ming, "An improved union model for continous speech recognition with partial duration corruption," in *ICASSP*, 2002.

[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification 2nd Edition*. Wiley, 2001.

[12] M. Siu and A. Chan, "Robust speech recognition against short-time noise," in *Proc ICSLP*, vol. 2, September 2002, pp. 1049–1052.

[13] P. Lamere, P. Kwok, W. Walker, E. Gouvilla, R. Singh, B. Raj, and P. Woolf, "Design of the cmu sphinx-4 decoder," in *Eurospeech*, 2003.

[14] R. G. Leonard, "A database for speaker-independent digit recognition," in *ICASSP*, 1984.

[15] S. J. Young, J. J. Odell, D. Olasen, and P. C. Woodland, *The HTK Book (for HTK Version 3.0)*, 1999.

[16] H. Hirsch and D. Pearse, “The aurora experimental framework for the performance evaluation of speech recognitions system under noisy conditions,” in *ASR 2000 Automatic Speech Recogntion : Challenges for the Next Millenium.*, 2000.

[17] P. Jancovic and J. Ming, “A probabilistic union model with automoatic order selection for noisy speech recognition,” *Journal of Acoustical Society of America*, vol. 110, September 2001.